

Aris Gkoulalas-Divanis
Grorios Loukides *Editors*

Medical Data Privacy Handbook

Medical Data Privacy Handbook

Aris Gkoulalas-Divanis • Grigorios Loukides
Editors

Medical Data Privacy Handbook



Springer

Editors

Aris Gkoulalas-Divanis
IBM Research - Ireland
Mulhuddart
Dublin, Ireland

Grigorios Loukides
Cardiff University
Cardiff, UK

ISBN 978-3-319-23632-2

DOI 10.1007/978-3-319-23633-9

ISBN 978-3-319-23633-9 (eBook)

Library of Congress Control Number: 2015947266

Springer Cham Heidelberg New York Dordrecht London
© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

*Dedicated to my parents, to Elena, and to the memory of my
grandmother Sophia*

-Aris Gkoulalas-Divanis

Dedicated to the memory of my grandmother

-Grigorios Loukides

Preface

The editors started working on medical data privacy in 2009, when they were postdoctoral researchers in the Health Information Privacy Laboratory, Department of Biomedical Informatics, Vanderbilt University. Their work on the topic involved understanding the privacy risks of medical data publishing and developing methods to prevent these risks. Protecting medical data privacy is a challenging problem, since a large volume of complex data must be protected in a setting that involves multiple parties (patients, physicians, carers, researchers, etc.). To address the problem, it is important to develop principled approaches that are specifically geared towards medical data. In addition, it is equally important to increase the awareness of all parties, involved in managing medical data, about privacy risks and approaches for achieving medical data privacy. Thus, the overarching aim of this book is to survey the field of medical data privacy and to present the state-of-the-art approaches to a wide audience.

The structure of the book closely follows the main categories of research works that have been undertaken to protect medical data privacy. Each such category is surveyed in a different part of the book, as follows. Part I is devoted to medical data *sharing*. Part II focuses on medical data privacy in *distributed and dynamic settings*. Following that, Part III examines privacy preservation in *emerging applications* featuring medical data, and Part IV discusses medical data privacy through *policy*, *data de-identification*, and *data governance*.

Privacy-preserving data sharing requires protecting the identity of patients and/or their sensitive information. For instance, attackers may use external data or background knowledge to learn patients' identity, even though attributes that directly identify patients (e.g., SSNs, phone numbers) have been removed. The problem has been studied extensively in the context of medical data, by the computer science, medical informatics, and statistics communities. However, there is no one-size-fits-all solution and various challenges remain. The purpose of Part I of this book is to survey the main research directions in the area of privacy-preserving medical data sharing and to present state-of-the-art approaches, including measures, algorithms, and software tools, that have been designed to solve this problem.

The protection of medical data privacy is particularly challenging, when multiple interrelated parties are involved. For example, medical data practitioners often need to link or exchange different parts of data about a patient, in the context of patient treatment. In addition, medical researchers or insurers may need to access patient information, according to the patient's privacy requirements. In this case, both the objectives of the parties accessing the data and the patient's requirements may change over time. Furthermore, data that are stored or processed in the cloud are vulnerable to a multitude of attacks, ranging from malicious access to intentional data modification. Part II of this book presents approaches focusing on privacy protection in such distributed and dynamic settings. These include approaches for linking data (record linkage), managing data access and patient consent, as well as exchanging health information. Furthermore, a comprehensive survey of privacy concerns and mitigation strategies for medical data in the cloud is presented.

Advances in medical devices and ubiquitous computing enable the collection and analysis of many complex data types, including genomic data, medical images, sensor data, biomedical signals, and health social network data. These data are valuable in a wide spectrum of emerging applications, either alone or in combination with data such as patient demographics and diagnosis codes, which are commonly found in Electronic Health Record (EHR) systems. For example, genomic studies have strong potential to lead to the discovery of effective, personalized drugs, and therapies. However, genomic data are extremely sensitive and must be privacy-protected. Part III of this book surveys privacy threats and solutions for all the aforementioned types of data that are central in emerging applications.

Parts I–III of this book focus on technical solutions that allow data owners (e.g., a healthcare institution) to effectively protect medical data privacy. On the other hand, Part IV focuses on the legal requirements for offering data privacy protection, as well as on the techniques and procedures that are required to satisfy this requirement. More specifically, this part examines key legal frameworks related to medical data privacy protection, as well as data de-identification and governance solutions, which are required to comply with these frameworks. A detailed presentation of the data protection legislation in the USA, EU, UK, and Canada is offered.

This book is primarily addressed to researchers and educators in the areas of computer science, statistics, and medical informatics who are interested in topics related to medical privacy. This book will also be a valuable resource to industry developers, as it explains the state-of-the-art algorithms for offering privacy. To ease understanding by nonexperts, the chapters contain a lot of background material, as well as many examples and citations to related literature. In addition, knowledge of medical informatics methods and terminology is not a prerequisite, and formalism was intentionally kept at a minimum. By discussing a wide range

of privacy techniques, providing in-depth coverage of the most important ones, and highlighting promising avenues for future research, this book also aims at attracting computer science and medical informatics students to this interesting field of research.

Dublin, Ireland
Cardiff, UK
July, 2015

Aris Gkoulalas-Divanis
Grigorios Loukides

Acknowledgements

We would like to thank all the authors, who have contributed chapters to this book, for their valuable contributions. This work would not have been possible without their efforts. A total of 63 authors who hold positions in leading academic institutions and industry, in Europe (France, Germany, Greece, Italy, Luxembourg, Switzerland, and UK), North America, Asia, Australia, and New Zealand, have contributed 29 chapters in this book, featuring more than 280 illustrations. We sincerely thank them for their hard work and the time they devoted to this effort.

In addition, we would like to express our deep gratitude to all the expert reviewers of the chapters for their constructive comments, which significantly helped towards improving the organization, readability, and overall quality of this handbook.

Last but not least, we are indebted to Susan Lagerstrom-Fife and Jennifer Malat from Springer, for their great guidance and advice in the preparation and completion of this handbook, as well as to the publication team at Springer for their valuable assistance in the editing process.

Contents

1	Introduction to Medical Data Privacy	1
	Aris Gkoulalas-Divanis and Grigoris Loukides	
1.1	Introduction	1
1.1.1	Privacy in Data Sharing	2
1.1.2	Privacy in Distributed and Dynamic Settings	3
1.1.3	Privacy for Emerging Applications	3
1.1.4	Privacy Through Policy, Data De-identification, and Data Governance.....	4
1.2	Part I: Privacy in Data Sharing	5
1.3	Part II: Privacy in Distributed and Dynamic Settings.....	8
1.4	Part III: Privacy for Emerging Applications	9
1.5	Part IV: Privacy Through Policy, Data De-identification, and Data Governance	11
1.6	Conclusion	13
	References	13

Part I Privacy in Data Sharing

2	A Survey of Anonymization Algorithms for Electronic Health Records	17
	Aris Gkoulalas-Divanis and Grigoris Loukides	
2.1	Introduction	17
2.2	Privacy Threats and Models	19
2.2.1	Privacy Threats	19
2.2.2	Privacy Models	19
2.3	Anonymization Algorithms	21
2.3.1	Algorithms Against Identity Disclosure	21
2.4	Directions for Future Research	29
2.5	Conclusion	31
	References	31

3 Differentially Private Histogram and Synthetic Data Publication	35
Haoran Li, Li Xiong, and Xiaoqian Jiang	
3.1 Introduction	35
3.2 Differential Privacy	36
3.2.1 Concept of Differential Privacy	36
3.2.2 Mechanisms of Achieving Differential Privacy	37
3.2.3 Composition Theorems	39
3.3 Relational Data	39
3.3.1 Problem Setting	39
3.3.2 Parametric Algorithms	42
3.3.3 Semi-parametric Algorithms	42
3.3.4 Non-parametric Algorithms	43
3.4 Transaction Data	48
3.4.1 Problem Setting	49
3.4.2 DiffPart	49
3.4.3 Private FIM Algorithms	50
3.4.4 PrivBasis	50
3.5 Stream Data	51
3.5.1 Problem Setting	51
3.5.2 Discrete Fourier Transform	52
3.5.3 FAST	52
3.5.4 w-Event Privacy	53
3.6 Challenges and Future Directions	54
3.6.1 Variety of Data Types	55
3.6.2 High Dimensionality	55
3.6.3 Correlated Constraints Among Attributes	55
3.6.4 Limitations of Differential Privacy	56
3.7 Conclusion	57
References	57
4 Evaluating the Utility of Differential Privacy: A Use Case	
Study of a Behavioral Science Dataset	59
Raquel Hill	
4.1 Introduction	59
4.2 Background	62
4.2.1 Syntactic Models: k -Anonymity	62
4.2.2 Differential Privacy: Definition	64
4.2.3 Applications	66
4.3 Methodology	67
4.3.1 Utility Measures	69
4.4 Results	70
4.4.1 Variable Distributions	71
4.4.2 Multivariate Logistic Regression	74
4.5 Discussion	79
4.6 Conclusion	80
References	80

5	SECRETA: A Tool for Anonymizing Relational, Transaction and RT-Datasets	83
	Giorgos Poulis, Aris Gkoulalas-Divanis, Grigorios Loukides, Spiros Skiadopoulos, and Christos Tryfonopoulos	
5.1	Introduction	84
5.2	Related Work	86
5.3	Overview of SECRETA	87
5.3.1	Frontend of SECRETA	87
5.3.2	Backend of SECRETA	93
5.3.3	Components	98
5.4	Using SECRETA	101
5.4.1	Preparing the Dataset	102
5.4.2	Using the Dataset Editor	103
5.4.3	The Hierarchy Editor	104
5.4.4	The Queries Workload Editor	104
5.4.5	Evaluating the Desired Method	105
5.4.6	Comparing Different Methods	106
5.5	Conclusion and Future Work	107
	References	108
6	Putting Statistical Disclosure Control into Practice: The ARX Data Anonymization Tool	111
	Fabian Prasser and Florian Kohlmayer	
6.1	Introduction	111
6.1.1	Background	112
6.1.2	Objectives and Outline	113
6.2	The ARX Data Anonymization Tool	114
6.2.1	Background	115
6.2.2	Overview	117
6.2.3	System Architecture	120
6.2.4	Application Programming Interface	123
6.2.5	Graphical User Interface	126
6.3	Implementation Details	133
6.3.1	Data Management	134
6.3.2	Pruning Strategies	136
6.3.3	Risk Analysis and Risk-Based Anonymization	138
6.4	Experimental Evaluation	139
6.5	Discussion	142
6.5.1	Comparison with Prior Work	142
6.5.2	Limitations and Future Work	144
6.5.3	Concluding Remarks	145
	References	145

7 Utility-Constrained Electronic Health Record Data Publishing Through Generalization and Disassociation	149
Grigorios Loukides, John Liagouris, Aris Gkoulalas-Divanis, and Manolis Terrovitis	
7.1 Introduction	150
7.1.1 Identity Disclosure	150
7.1.2 Utility-Constrained Approach	152
7.1.3 Chapter Organization	154
7.2 Preliminaries	155
7.3 Generalization and Disassociation	156
7.4 Specification of Utility Constraints	159
7.4.1 Defining and Satisfying Utility Constraints	159
7.4.2 Types of Utility Constraints for ICD Codes	162
7.5 Utility-Constrained Anonymization Algorithms	163
7.5.1 Clustering-Based Anonymizer (CBA)	164
7.5.2 DISassociation Algorithm (DIS)	165
7.5.3 Comparing the CBA and DIS Algorithms	169
7.6 Future Directions	174
7.6.1 Different Forms of Utility Constraints	174
7.6.2 Different Approaches to Guaranteeing Data Utility	175
7.7 Conclusion	176
References	176
8 Methods to Mitigate Risk of Composition Attack in Independent Data Publications	179
Jiuyong Li, Sarowar A. Sattar, Muzammil M. Baig, Jixue Liu, Raymond Heatherly, Qiang Tang, and Bradley Malin	
8.1 Introduction	180
8.2 Composition Attack and Multiple Data Publications	181
8.2.1 Composition Attack	181
8.2.2 Multiple Coordinated Data Publications	183
8.2.3 Multiple Independent Data Publications	183
8.3 Risk Mitigation Through Randomization	185
8.4 Risk Mitigation Through Generalization	187
8.5 An Experimental Comparison	189
8.5.1 Data and Setting	190
8.5.2 Reduction of Risk of Composition Attacks	190
8.5.3 Comparison of Utility of the Two Methods	192
8.6 Risk Mitigation Through Mixed Publications	193
8.7 Conclusion	196
References	198

9 Statistical Disclosure Limitation for Health Data: A Statistical Agency Perspective	201
Natalie Shlomo	
9.1 Introduction	201
9.2 Statistical Disclosure Limitation for Microdata from Social Surveys	203
9.2.1 Disclosure Risk Assessment	204
9.2.2 Statistical Disclosure Limitation Methods	207
9.2.3 Information Loss Measures	211
9.3 Statistical Disclosure Limitation for Frequency Tables	213
9.3.1 Disclosure Risk in Whole Population Tabular Outputs ...	213
9.3.2 Disclosure Risk and Information Loss Measures Based on Information Theory	214
9.3.3 Statistical Disclosure Limitation Methods	217
9.4 Differential Privacy in Survey Sampling and Perturbation	219
9.5 Future Outlook for Releasing Statistical Data	222
9.5.1 Safe Data Enclaves and Remote Access	223
9.5.2 Web-Based Applications	224
9.5.3 Synthetic Data	226
9.6 Conclusion	228
References	228

Part II Privacy in Distributed and Dynamic Settings

10 A Review of Privacy Preserving Mechanisms for Record Linkage ...	233
Luca Bonomi, Liyue Fan, and Li Xiong	
10.1 Introduction	233
10.2 Overview of Privacy Preserving Record Linkage.....	236
10.2.1 The PPRL Model	236
10.2.2 Taxonomy of Presented Techniques	238
10.3 Secure Transformations	244
10.3.1 Attribute Suppression and Generalization Methods	245
10.3.2 N-Grams Methods	246
10.3.3 Embedding Methods	248
10.3.4 Phonetic Encoding Methods	250
10.4 Secure Multi-Party Computation	251
10.4.1 Commutative Encryption Based Protocols	251
10.4.2 Homomorphic Encryption Based Protocols	252
10.4.3 Secure Scalar Product Protocols	254
10.5 Hybrid Approaches	256
10.5.1 Standard Blocking	257
10.5.2 Sorted Neighborhood Approach	258
10.5.3 Mapping	259
10.5.4 Clustering	259
10.6 Challenges and Future Research Directions	261

10.7 Conclusion	262
References	262
11 Application of Privacy-Preserving Techniques in Operational Record Linkage Centres	267
James H. Boyd, Sean M. Randall, and Anna M. Ferrante	
11.1 Introduction	267
11.1.1 Record Linkage Research Infrastructure	268
11.1.2 Privacy Challenges in Health Record Linkage	270
11.2 Data Governance	271
11.2.1 Legal Obligations	272
11.2.2 Information Governance	272
11.2.3 Separation of Data and Functions	273
11.2.4 Application and Approval Process	273
11.2.5 Information Security	274
11.3 Operational Models and Data Flows	274
11.3.1 Centralized Model	275
11.3.2 Separated Models	276
11.3.3 A Technique to Avoid Data Collusion	278
11.4 Privacy Preserving Methods	278
11.4.1 Privacy Preserving Models	279
11.4.2 Techniques for Privacy Preserving Linkage	279
11.4.3 Requirements of a Privacy Preserving Linkage Technique for Operational Linkage Centres	282
11.5 Conclusion	285
References	285
12 Privacy Considerations for Health Information Exchanges	289
Dalvin Hill, Joseph Walker, and John Hale	
12.1 Introduction	289
12.2 Health Information Exchanges	290
12.2.1 HIE Actors and Systems	290
12.2.2 HIE Models	293
12.2.3 HIPAA, HITECH and HIE Privacy Governance	294
12.3 Privacy Issues with HIEs	295
12.3.1 Patient Expectations and Concerns	296
12.3.2 Tension Between Functionality, Security and Privacy	297
12.3.3 Data Stewardship and Ownership	297
12.4 Principles and Practice of Privacy for HIEs	298
12.4.1 Guiding Principles	298
12.4.2 HIE Privacy in Practice	300
12.5 Emerging Issues	305
12.5.1 Big Data	305
12.5.2 m-Health and Telemedicine	306
12.5.3 Medical Devices	307

12.6 Conclusion	308
References	308
13 Managing Access Control in Collaborative Processes for Healthcare Applications	313
Xuan Hung Le and Dongwen Wang	
13.1 Introduction	314
13.2 Related Works	314
13.3 An Illustrative Example: New York State HIV Clinical Education Initiative	316
13.4 Development of the Enhanced RBAC Model	318
13.4.1 Overview of the Enhanced RBAC Model	319
13.4.2 Support Team Collaboration: Bridging Entities and Contributing Attributes	320
13.4.3 Extending Access Permissions to Include Workflow Contexts	322
13.4.4 Role-Based Access Delegation Targeting on Specific Objects: Providing Flexibility for Access Control in Collaborative Processes	322
13.4.5 Integration of Multiple Representation Elements for Definition of Universal Constraints	324
13.4.6 Case Studies to Encode Access Policies for CEI	326
13.5 System Framework for Implementation of Enhanced RBAC	329
13.5.1 System Architecture	330
13.5.2 Encoding of Access Policies	331
13.5.3 Interpretation of Access Control Policies	333
13.5.4 Application Layer	334
13.5.5 Demonstration Tool	334
13.6 Evaluation of the Enhanced RBAC Model	335
13.6.1 Selection of Study Cases	336
13.6.2 Access Permissions Computed with the Enhanced RBAC Model and the CEIAdmin System	339
13.6.3 Comparison Between the Enhanced RBAC Model and the CEIAdmin System	340
13.6.4 Development of the Gold-Standard	340
13.6.5 Measuring Effectiveness Based on Gold-Standard	342
13.6.6 Results	344
13.7 Discussion	345
13.7.1 Features of the Enhanced RBAC Model	345
13.7.2 System Framework for Implementation	349
13.7.3 Evaluation	350
13.7.4 Limitations	353
13.8 Conclusion	354
References	355

14	Automating Consent Management Lifecycle for Electronic Healthcare Systems	361
	Muhammad Rizwan Asghar and Giovanni Russello	
14.1	Introduction	361
14.2	Legal Background	363
14.2.1	Legal Framework for Consent	363
14.2.2	Consent in Healthcare Systems	365
14.2.3	Consent Limitations	366
14.3	A Case Study	368
14.4	Overview of Teleo-Reactive Policies	369
14.4.1	TR Policy Representation	369
14.4.2	TR Policy Evaluation	370
14.5	The ACTORS Approach	371
14.5.1	Authorisation Policies	373
14.5.2	Policy Templates	374
14.5.3	TR Policies	375
14.6	Managing Consent in Healthcare Scenarios	376
14.7	Related Work	382
14.8	Conclusion and Future Work	384
	References	385
15	e-Health Cloud: Privacy Concerns and Mitigation Strategies	389
	Assad Abbas and Samee U. Khan	
15.1	Introduction	389
15.2	An Overview of the e-Health Cloud	391
15.2.1	e-Health Cloud Benefits and Opportunities	391
15.2.2	Deployment Models for Cloud Based e-Health Systems	393
15.2.3	Threats to Health Data Privacy in the Cloud	394
15.2.4	Essential Requirements for Privacy Protection	397
15.2.5	User/Patient Driven Privacy Protection Requirements	399
15.2.6	Adversarial Models in the e-Health Cloud	399
15.3	Privacy Protection Strategies Employed in e-Health Cloud	400
15.3.1	Approaches to Protect Confidentiality in the e-Health Cloud	400
15.3.2	Approaches to Maintain Data Integrity in the e-Health Cloud	402
15.3.3	Approaches to Offer Collusion Resistance in the e-Health Cloud	406
15.3.4	Approaches to Maintain Anonymity in the e-Health Cloud	407
15.3.5	Approaches to Offer Authenticity in the e-Health Cloud	410
15.3.6	Approaches to Maintain Unlinkability in the e-Health Cloud	412
15.4	Discussion and Open Research Issues	416

15.5 Conclusion	417
References	418
Part III Privacy for Emerging Applications	
16 Preserving Genome Privacy in Research Studies	425
Shuang Wang, Xiaoqian Jiang, Dov Fox, and Lucila Ohno-Machado	
16.1 Introduction	426
16.2 Policies, Legal Regulation and Ethical Principles of Genome Privacy	427
16.2.1 NIH Policies for Genomic Data Sharing	427
16.2.2 U.S. Legal Regulations for Genomic Data	430
16.2.3 Ethical Principles for Genome Privacy	432
16.2.4 Summary	433
16.3 Information Technology for Genome Privacy	433
16.3.1 Genome Privacy Risks	434
16.3.2 Genome Privacy Protection Technologies	434
16.3.3 Community Efforts on Genome Privacy Protection.....	436
16.4 Conclusion	437
References	438
17 Private Genome Data Dissemination	443
Noman Mohammed, Shuang Wang, Rui Chen, and Xiaoqian Jiang	
17.1 Introduction	443
17.2 Literature Review	445
17.2.1 Privacy Attacks and Current Practices	445
17.2.2 Privacy Preserving Techniques.....	446
17.3 Problem Statement	447
17.3.1 Privacy Protection Model	448
17.3.2 Privacy Attack Model	448
17.3.3 Utility Criteria	449
17.4 Genomic Data Anonymization	449
17.4.1 Anonymization Algorithm	449
17.4.2 Privacy Analysis	453
17.4.3 Computational Complexity	453
17.5 Experimental Results.....	454
17.6 Conclusion	458
References	459
18 Threats and Solutions for Genomic Data Privacy	463
Erman Ayday and Jean-Pierre Hubaux	
18.1 Threats for Genomic Privacy	463
18.1.1 Kin Genomic Privacy	465

18.2	Solutions for Genomic Privacy	470
18.2.1	Privacy-Preserving Management of Raw Genomic Data.....	470
18.2.2	Private Use of Genomic Data in Personalized Medicine	472
18.2.3	Private Use of Genomic Data in Research	477
18.2.4	Coping with Weak Passwords for the Protection of Genomic Data.....	481
18.2.5	Protecting Kin Genomic Privacy.....	484
18.3	Future Research Directions	487
18.4	Conclusion	490
	References	490
19	Encryption and Watermarking for medical Image Protection	493
	Dalel Bouslimi and Gouenou Coatrieux	
19.1	Introduction.....	493
19.2	Security Needs for Medical Data	495
19.2.1	General Framework.....	495
19.2.2	Refining Security Needs in an Applicative Context: Telemedicine Applications as Illustrative Example	497
19.3	Encryption Mechanisms: An <i>A Priori</i> Protection.....	498
19.3.1	Symmetric/Asymmetric Cryptosystems & DICOM	498
19.3.2	Block Cipher/Stream Cipher Algorithms.....	499
19.4	Watermarking: An <i>A Posteriori</i> Protection Mechanism	503
19.4.1	Principles, Properties and Applications	503
19.4.2	Watermarking Medical Images	506
19.5	Combining Encryption with Watermarking	512
19.5.1	Continuous Protection with Various Security Objectives: A State of the Art	512
19.5.2	A Joint Watermarking-Encryption (JWE) Approach	516
19.6	Conclusion	521
	References	521
20	Privacy Considerations and Techniques for Neuroimages	527
	Nakeisha Schimke and John Hale	
20.1	Introduction	527
20.2	Neuroimage Data	529
20.3	Privacy Risks with Medical Images	530
20.3.1	Neuroimage Privacy Threat Scenarios	530
20.3.2	Volume Rendering and Facial Recognition.....	532
20.3.3	Re-identification Using Structural MRI	534
20.4	Privacy Preservation Techniques for Medical Images	535
20.4.1	De-Identification Techniques	535
20.4.2	Privacy in Neuroimage Archives and Collaboration Initiatives	543

Contents	xxiii
20.5 Conclusion	544
References	544
21 Data Privacy Issues with RFID in Healthcare	549
Peter J. Hawrylak and John Hale	
21.1 Introduction	549
21.1.1 RFID as a Technology	550
21.2 Dimensions of Privacy in Medicine	553
21.3 RFID in Medicine	556
21.3.1 Inventory Tracking	556
21.3.2 Tracking People	556
21.3.3 Device Management	557
21.4 Issues and Risks	558
21.5 Solutions	562
21.6 Conclusion	563
References	564
22 Privacy Preserving Classification of ECG Signals in Mobile e-Health Applications	569
Riccardo Lazzeretti and Mauro Barni	
22.1 Introduction	569
22.2 Plain Protocol	572
22.2.1 Classification Results	575
22.3 Cryptographic Primitives	575
22.3.1 Homomorphic Encryption	576
22.3.2 Oblivious Transfer	577
22.3.3 Garbled Circuits	578
22.3.4 Hybrid Protocols	579
22.4 Privacy Preserving Linear Branching Program	580
22.4.1 Linear Branching Programs (LBP)	580
22.4.2 ECG Classification Through LBP and Quadratic Discriminant Functions	584
22.4.3 ECG Classification Through LBP and Linear Discriminant Functions	586
22.4.4 Complexity Analysis	587
22.5 Privacy Preserving Classification by Using Neural Network	590
22.5.1 Neural Network Design	590
22.5.2 Quantized Neural Network Classifier	593
22.5.3 Privacy-Preserving GC-Based NN Classifier	595
22.5.4 Privacy-Preserving Hybrid NN Classifier	597
22.5.5 Comparison with the LBP Solution	598
22.6 Privacy Preserving Quality Evaluation	599
22.6.1 SNR Evaluation in the Encrypted Domain	599
22.6.2 SNR-Based Quality Evaluation	603
22.7 Conclusion	608
References	609

23 Strengthening Privacy in Healthcare Social Networks	613
Maria Bertsimas, Iraklis Varlamis, and Panagiotis Rizomiliotis	
23.1 Introduction	613
23.2 Social Networks	615
23.2.1 On-line Social Networks	615
23.2.2 Healthcare Social Networks	616
23.3 Privacy	618
23.3.1 Background	618
23.3.2 Personal and Sensitive Data	619
23.3.3 Privacy Principles	621
23.3.4 Privacy Threats	622
23.4 Privacy Requirements for HSNs	627
23.4.1 Privacy as System Requirement	627
23.5 Enhancing Privacy in OSNs and HSNs	628
23.6 On-line Social Networks in the Healthcare Domain	631
23.6.1 Advice Seeking Networks	632
23.6.2 Patient Communities	632
23.6.3 Professional Networks	633
23.7 Conclusion	633
References	634

Part IV Privacy Through Policy, Data De-identification, and Data Governance

24 Privacy Law, Data Sharing Policies, and Medical Data: A Comparative Perspective	639
Edward S. Dove and Mark Phillips	
24.1 Introduction	639
24.2 Overview of Data Privacy Legal Frameworks	642
24.3 Data Privacy Laws and Guidelines	648
24.3.1 The OECD Privacy Guidelines	648
24.3.2 The Council of Europe Convention 108	650
24.3.3 The European Union Data Protection Directive 95/46	652
24.3.4 UK Data Protection Act 1998	656
24.3.5 Canadian Privacy Legislation	658
24.3.6 The HIPAA Privacy Rule	659
24.4 Data Sharing Policies	664
24.4.1 US National Institutes of Health	665
24.4.2 Canadian Data Sharing Policies	666
24.4.3 Wellcome Trust (UK)	669
24.5 Towards Better Calibration of Biomedical Research, Health Service Delivery, and Privacy Protection	671
24.6 Conclusion	674
References	674

25 HIPAA and Human Error: The Role of Enhanced Situation Awareness in Protecting Health Information	679
Divakaran Liginla	
25.1 Introduction	679
25.2 HIPAA, Privacy Breaches, and Related Costs	682
25.3 Situation Awareness and Privacy Protection	685
25.3.1 Definition of Situation Awareness	685
25.3.2 Linking Situation Awareness to Privacy Breaches	686
25.3.3 SA and HIPAA Privacy Breaches	688
25.4 Discussion and Conclusion	693
References	695
26 De-identification of Unstructured Clinical Data for Patient Privacy Protection	697
Stephane M. Meystre	
26.1 Introduction	697
26.2 Origins and Definition of Text De-identification	698
26.3 Methods Applied for Text De-identification	701
26.4 Clinical Text De-identification Application Examples	704
26.4.1 Physionet Deid	704
26.4.2 MIST (MITRE Identification Scrubber Toolkit)	705
26.4.3 VHA Best-of-Breed Clinical Text De-identification System	706
26.5 Why Not Anonymize Clinical Text?	708
26.6 U.S. Veterans Health Administration Clinical Text De-identification Efforts	709
26.7 Conclusion	713
References	714
27 Challenges in Synthesizing Surrogate PHI in Narrative EMRs	717
Amber Stubbs, Özlem Uzuner, Christopher Kotfila, Ira Goldstein, and Peter Szolovits	
27.1 Introduction	717
27.2 Related Work	719
27.3 PHI Categories	722
27.4 Data	724
27.5 Strategies and Difficulties in Surrogate PHI Generation	725
27.5.1 HIPAA Category 1: Names	726
27.5.2 HIPAA Category 2: Locations	728
27.5.3 HIPAA Category 3: Dates and Ages	729
27.5.4 HIPAA Category 18: Other Potential Identifiers	731
27.6 Errors Introduced by Surrogate PHI	732
27.7 Relationship Between De-identification and Surrogate Generation	732
27.8 Conclusion	733
References	734

28 Building on Principles: The Case for Comprehensive, Proportionate Governance of Data Access.....	737
Kimberlyn M. McGrail, Kaitlyn Gutteridge, and Nancy L. Meagher	
28.1 Introduction.....	737
28.2 Current Approaches to Data Access Governance.....	739
28.2.1 Existing Norms for Data Access Governance	739
28.2.2 The Preeminence of “Consent or Anonymize” as Approaches to Data Access Governance	740
28.2.3 Existing Data Access Governance in Practice.....	743
28.3 The Evolution of Data and Implications for Data Access Governance.....	744
28.3.1 Big Data.....	744
28.3.2 Open Data.....	745
28.3.3 The Ubiquity of Collection of Personal Information	745
28.3.4 The Limits of Existing Approaches to Data Access Governance	746
28.4 A Comprehensive Model for Governance: Proportionate and Principled.....	747
28.4.1 Proportionality	747
28.4.2 Principle-Based Regulation	748
28.4.3 Case Studies Using Proportionate and Principled Access	749
28.5 Building on the Present: A Flexible, Governance Framework	752
28.5.1 Science	754
28.5.2 Approach	754
28.5.3 Data	755
28.5.4 People	755
28.5.5 Environment	755
28.5.6 Interest	756
28.5.7 Translating Risk Assessment to Review Requirements...	756
28.5.8 Adjudication Scenarios	757
28.6 Conclusion	759
References.....	760
29 Epilogue	765
Aris Gkoulalas-Divanis and Grigoris Loukides	
29.1 Introduction	765
29.2 Topics and Directions in Privacy Preserving Data Sharing.....	766
29.3 Topics and Directions in Privacy Preservation for Distributed and Dynamic Settings	768
29.4 Topics and Directions in Privacy Preservation for Emerging Applications	769
29.5 Topics and Directions in Privacy Preservation Through Policy, Data De-identification, and Data Governance	771

Contents	xxvii
29.6 Conclusion	772
References	772
About the Authors	775
Glossary	815
Index	827

List of Figures

Fig. 3.1	Example: released cell histogram (<i>left</i>) and subcube histogram (<i>right</i>), and N_i is a random Laplace noise (see Sect. 3.2 for Laplace mechanism)	40
Fig. 3.2	Generate synthetic data via parametric methods	41
Fig. 3.3	Generate synthetic data via non-parametric methods	41
Fig. 3.4	Generate synthetic data via semi-parametric methods	41
Fig. 3.5	DExample of private quadtree: noisy counts (<i>inside boxes</i>) are released; actual counts, although depicted, are not released. Query Q (<i>dotted red rectangle</i>) could be answered by adding noisy counts of marked nodes (Color figure online) [6]	45
Fig. 3.6	Taxonomy tree of attributes [29]	48
Fig. 3.7	Tree for partitioning records [29]	48
Fig. 3.8	A context-free taxonomy tree of the sample data in Table 3.1 [5]	49
Fig. 3.9	The partitioning process of Fig. 3.1 [5]	50
Fig. 3.10	The FAST framework [16]	53
Fig. 4.1	Excerpt from doctor's notes	60
Fig. 4.2	Experiment flow chart	67
Fig. 4.3	Histogram of ages from original data (<i>left</i>) and using k-d tree algorithm with $\epsilon = 2.0$, ET = 0.677 (<i>right</i>)	72
Fig. 4.4	Histogram of genders from original data (<i>left</i>) and using cell-based algorithm with $\epsilon = 2.0$ (<i>right</i>)	72
Fig. 4.5	Proportion of variable counts vs. ϵ for all algorithms (for the first reduced dataset)	73
Fig. 4.6	Proportion of variable counts vs. ϵ for all algorithms (for the second reduced dataset)	73
Fig. 4.7	Proportion of variable counts preserved vs. ϵ for k-d tree (for MART_rs1)	74

Fig. 4.8	Proportion of variable counts preserved vs. ϵ for k-d tree (for MART_rs2)	75
Fig. 4.9	Effect size versus ϵ for RS1 cell-based runs that were similar	75
Fig. 4.10	Logistic results for MART_final for k-d tree, effect size and proportion of good runs versus the DP ϵ parameter	76
Fig. 4.11	Logistic results for MART_rs1 for k-d tree, proportion of good runs versus the DP ϵ parameter	77
Fig. 4.12	Logistic results for MART_rs1 for k-d tree, effect size and proportion of good runs versus the DP ϵ parameter for entropy_threshold = 1.0	77
Fig. 4.13	Logistic results for MART_rs2, proportion of good runs versus the DP ϵ parameter	78
Fig. 4.14	Logistic results for MART_rs2 for k-d tree, effect size and proportion of good runs versus the DP ϵ parameter for entropy_threshold = 1.0	78
Fig. 5.1	System architecture of SECRETA	88
Fig. 5.2	The main screen of SECRETA	89
Fig. 5.3	Automatic creation of hierarchies. (a) Selecting the number of splits per level of the hierarchy and (b) Displaying the produced hierarchy	89
Fig. 5.4	The evaluation mode: method evaluation screen of SECRETA	90
Fig. 5.5	The comparison mode: methods comparison screen of SECRETA	91
Fig. 5.6	The experimentation interface selector	91
Fig. 5.7	Plots for (a) the original dataset, (b) varying parameters execution, and (c) the comparison mode	92
Fig. 5.8	An example of a hierarchy tree	94
Fig. 5.9	The dataset editor	103
Fig. 5.10	Frequency plots of the original dataset	103
Fig. 5.11	The hierarchy specification area	104
Fig. 5.12	Method parameters setup	105
Fig. 5.13	A messagebox with the results summary	106
Fig. 5.14	The data output area	106
Fig. 5.15	The plotting area	106
Fig. 5.16	The configurations editor	107
Fig. 6.1	Example cancer dataset: types of attributes and types of disclosure	115
Fig. 6.2	Generalization hierarchies for attributes age and gender	116
Fig. 6.3	Example search space	117
Fig. 6.4	High-level architecture of the ARX system	120
Fig. 6.5	Overview of the most important classes in ARX's core	121
Fig. 6.6	Overview of the most important classes in ARX's application programming interface	122
Fig. 6.7	Anonymization process implemented in ARX's GUI	127

Fig. 6.8	The ARX configuration perspective	128
Fig. 6.9	Wizard for creating a generalization hierarchy with intervals.....	129
Fig. 6.10	The ARX exploration perspective	130
Fig. 6.11	The ARX utility evaluation perspective	132
Fig. 6.12	The ARX risk analysis perspective	133
Fig. 6.13	Example of how data is encoded and transformed in ARX	134
Fig. 6.14	Example of how data snapshots are represented in ARX.....	135
Fig. 8.1	The average accuracy of the composition attack on the <i>Salary</i> and <i>Occupation</i> datasets	191
Fig. 8.2	The average query errors of the <i>Salary</i> and <i>Occupation</i> datasets with different methods	192
Fig. 8.3	Distance between the original dataset, the output of dLink, and several privacy budgets of differential privacy ($\epsilon = 0.01, 0.05, 0.1$)	193
Fig. 8.4	An illustration of the mixed publication model	194
Fig. 8.5	An example of the mixed publication	195
Fig. 9.1	Confidential residual plot from a regression analysis on receipts for the Sugar Canes dataset. (a) Residuals by fitted values. (b) Normal QQ plot of residuals.....	226
Fig. 9.2	Univariate analysis of receipts for the Sugar Canes dataset	227
Fig. 10.1	The privacy-preserving record linkage (PPRL) model	237
Fig. 10.2	Bloom Filter representation for the names SMITH and SMYTH using 2-g. The map is obtained using one hash function and in total there are ten bits in A, and 11 bits in B set to 1. Only eight bits are shared among the Bloom filters, therefore the similarity measure between the original strings approximated with the Dice coefficient is $\frac{2 \cdot 8}{(10+11)} \approx 0.762$ (example from Schnell et al. [45])	246
Fig. 10.3	Example of composite Bloom filter representation from Durham et al. [13]. (a) Transformation process. (b) Composite bloom filter	247
Fig. 10.4	Embedding example for the names SMITH and SMYT with an embedding base formed by the sets S_1, S_2, S_3, S_4 of randomly generated strings	249
Fig. 10.5	Example of bitwise encryption by Kuzu et al. [33]	253
Fig. 10.6	Performing blocking on datasets: (a) Original datasets T and V ; (b) Block decomposition using hyper-rectangles for T ; (c) Block perturbation for T ; (d) Block perturbation for V . The candidate matching pairs tested in the SMC part are limited to the pair of overlapping blocks: $(T_1, V_1), (T_1, V_2), \dots, (T_5, V_5)$ (example from Inan et al. [24])	258

Fig. 10.7	Private blocking via clustering (example from [28])	260
Fig. 11.1	A centralized model: data providers give full datasets to the linkage unit, who link and then pass on the data to the researcher	275
Fig. 11.2	The data provider splits the data, sending the personal identifiers to the linkage unit and the clinical content to the client services team. The linkage unit then provides the linkage map to the client services team who join it to content data to create datasets for research and analysis	277
Fig. 11.3	In the absence of a repository of clinical data, this is supplied to the researcher by the data provider	278
Fig. 11.4	Basic data flow of three party privacy preserving linkage	280
Fig. 11.5	Numerous protocols attempt to reduce the variability between records of the same person, while maintaining variability between records belonging to different people	281
Fig. 11.6	Creating a statistical linkage key	281
Fig. 11.7	First and last name are phonetically encoded and concatenated with date of birth and sex, which is then hashed to form the Swiss Anonymous Linkage Code	282
Fig. 12.1	HIE actors and systems	292
Fig. 12.2	HIE models: (a) centralized, (b) decentralized, and (c) hybrid....	293
Fig. 12.3	HIPAA covered entities and business associates	295
Fig. 13.1	Workflow of a CEI training session (reprinted from [51], with permission from Elsevier)	318
Fig. 13.2	Enhanced RBAC model with universal constraints, workflow in permissions, and domain ontologies (reprinted from [51], with permission from Elsevier)	320
Fig. 13.3	Collaboration model with bridging entity and contributing attributes (reprinted from [51], with permission from Elsevier).....	321
Fig. 13.4	System architecture (reprinted from [49]).....	331
Fig. 13.5	Three-level access control policy encoding in Protégé (reprinted from [49])	332
Fig. 13.6	An example of access policy for CEI (reprinted from [49]); (a) Access policy in first-order predicate logic, (b) Access policy in Protege SWRL	333
Fig. 13.7	A screenshot of the demo tool showing CEI access management (reprinted from [49])	335
Fig. 13.8	Overall design of the evaluation study (reprinted from [52], with permission from Elsevier)	336

Fig. 13.9	Mappings of CEI Centers, system roles, and users (reprinted from [52], with permission from Elsevier)	337
Fig. 13.10	A screenshot of the online tool used by the judges to build the gold-standard (reprinted from [52], with permission from Elsevier).....	343
Fig. 13.11	Mapping the enhanced RBAC framework to XACML (reprinted from [49])	350
Fig. 14.1	A layout of TR policies	369
Fig. 14.2	An example of a TR policy	370
Fig. 14.3	The ACTORS architecture for managing consent lifecycle	372
Fig. 14.4	An example of an authorisation policy	374
Fig. 14.5	An example of a policy template	375
Fig. 14.6	A TR policy for managing authorisation policy for providing consent to a GP	377
Fig. 14.7	A policy template for generating an authorisation policy for providing consent to a GP	377
Fig. 14.8	An authorisation policy for providing consent to a GP	378
Fig. 14.9	A TR policy for providing consent to a specialist	379
Fig. 14.10	A policy template for generating an authorisation policy for providing consent to a cardiologist.....	380
Fig. 14.11	An authorisation policy for providing consent to a cardiologist ..	380
Fig. 14.12	A policy template for generating an authorisation policy for providing consent to the emergency response team	381
Fig. 14.13	An authorisation policy for providing consent to the emergency response team	381
Fig. 15.1	Distinction among the EMR, PHR, and EHR	392
Fig. 15.2	An illustration of a private cloud in context of e-Health	394
Fig. 15.3	An illustration of a public cloud in context of e-Health	395
Fig. 15.4	An illustration of a hybrid cloud in context of e-Health	395
Fig. 15.5	Taxonomy of essential privacy requirements and patient-driven requirements	397
Fig. 16.1	HHS's new rules to address the risks of de-identified data that can be re-identified	431
Fig. 16.2	Privacy protection and number of released independent single nucleotide polymorphisms (SNPs) base on the report in [63].....	434

Fig. 16.3	Illustration of Homer's attacks, where $ P_j - R_j $ and $ P_j - M_j $ measure how the person's allele frequency P_j differs from the allele frequencies of the reference population and the mixture, respectively. By sampling a large number of N SNPs, the distance measure $D(P_j)$ will follow a normal distribution due to the central limit theorem. Then, a one-sample t-test for this individual over all sampled SNPs can be used to verify the hypothesis that an individual is in the mixture ($T(P) > 0$)	435
Fig. 17.1	Taxonomy tree of blocks	451
Fig. 17.2	Tree for partitioning records	452
Fig. 17.3	Boxplots of data utility of chr2 and chr10 data with different p-values	456
Fig. 17.4	Privacy risk of chr2 and chr10 data. The <i>star</i> and <i>diamond</i> markers represent the test value of a specific individual in the case (<i>left</i>) or test (<i>right</i>) group, respectively. The <i>horizontal line</i> indicates the 0.95 confidence level for identifying case individuals that are estimated based on the test statistic values of test individuals	457
Fig. 17.5	Comparison of data utility and privacy risk for chr2 and chr10 data with different privacy budget	458
Fig. 18.1	Overview of the proposed framework to quantify kin genomic privacy [23]. Each vector X^i ($i \in \{1, \dots, n\}$) includes the set of SNPs for an individual in the targeted family. Furthermore, each letter pair in X^i represents a SNP x_j^i ; and for simplicity, each SNP x_j^i can be represented using $\{BB, Bb, bb\}$ (or $\{0, 1, 2\}$). Once the health privacy is quantified, the family should ideally decide whether to reveal less or more of their genomic information through the genomic-privacy preserving mechanism (GPPM)	466
Fig. 18.2	Family tree of <i>CEPH/Utah Pedigree 1463</i> consisting of the 11 family members that were considered. The notations <i>GP</i> , <i>P</i> , and <i>C</i> stand for “grandparent”, “parent”, and “child”, respectively. Also, the symbols σ and φ represent the male and female family members, respectively	468

Fig. 18.3	Evolution of the genomic privacy of the parent (P5), with and without considering LD. For each family member, we reveal 50 randomly picked SNPs (among 100 SNPs in S), starting from the most distant family members, and the x-axis represents the exact sequence of this disclosure. Note that $x = 0$ represents the prior distribution, when no genomic data is revealed	469
Fig. 18.4	Connections between the parties in the proposed protocol for privacy-preserving management of raw genomic data [4]	472
Fig. 18.5	Parts to be masked in the short reads for out-of-range content	473
Fig. 18.6	Parts to be masked in a short read based on patient's consent. The patient does not give consent to reveal the dark parts of the short read.....	473
Fig. 18.7	Proposed privacy-preserving disease susceptibility test (PDS) [6]	476
Fig. 18.8	Proposed system model for the privacy-preserving computation of the disease risk [5]	478
Fig. 18.9	System model for private use of genomic data in research setting [38]: participants (P), certified institution (CI), storage and processing unit (SPU), and medial units (MU)	481
Fig. 18.10	GenoGuard protocol [38]. A patient provides his biological sample to the CI, and chooses a password for honey encryption. The CI does the sequencing, encoding and password-based encryption, and then sends the ciphertext to the biobank. During a retrieval, a user (e.g., the patient or his doctor) requests for the ciphertext, decrypts it and finally decodes it to get the original sequence	483
Fig. 18.11	General protection framework. The GPPM [24] takes as inputs (i) the privacy levels of all family members, (ii) the genome of the donor, (iii) the privacy preferences of the family members, and (iv) the research utility. First, correlations between the SNPs (LD) is not considered in order to use combinatorial optimization. Note that we go only once through this box. Then, LD is used and a fine-tuning algorithm is used to cope with non-linear constraints. The algorithm outputs the set of SNPs that the donor can disclose	485
Fig. 19.1	A cryptosystem	498
Fig. 19.2	General scheme of the AES algorithm	501

Fig. 19.3	AES encryption in CBC mode. B_i , B_i^e and K_e denote the plaintext block, the encrypted block and the encryption key, respectively. iv is a random initialization vector	502
Fig. 19.4	Encryption/decryption processes of a stream cipher algorithm ...	502
Fig. 19.5	The principle of watermarking chain	504
Fig. 19.6	Example of two codebooks' cells in the mono-dimensional space (i.e. x is a scalar value) considering an uniform quantization of quantization step Δ . Symbols \circ and \times denote cells' centers that encode 0 and 1, respectively. $d = \Delta/2$ establishes the measure of robustness to signal perturbations	509
Fig. 19.7	Insertion of a binary message using AQIM. X_w represents the vector after the insertion of a bit equals to "1" within a host signal X associated to a vector in the N -dimensional space if N pixels constitute X . Δ is the quantization step, and circles and crosses represent the centers of the cells that encode the bits "0" and "1", respectively	510
Fig. 19.8	Basic principle of the histogram shifting modulation: (a) original histogram, and (b) histogram of the watermarked data	511
Fig. 19.9	General system architecture of a JWE algorithm. M_e , M_s , K_e , K_{ws} and K_{we} are the message available in the encrypted domain, the message available in the spatial domain, the encryption and the watermarking keys in the spatial and the encrypted domain, respectively	517
Fig. 19.10	Examples of the images used for evaluation (using AES): (a) original PET image, (b) joint watermarked/ciphered image, (c) deciphered watermarked image, and (d) difference between the original image and the decrypted watermarked image	520
Fig. 20.1	Re-linkage using an imaging database	531
Fig. 20.2	Volume rendering from 3D Slicer	532
Fig. 20.3	Skull Stripping: 3dSkullstrip in AFNI (<i>left</i>); BET in FSL (<i>middle</i>); HWA in Freesurfer (<i>right</i>) [41]	536
Fig. 20.4	Defacing: Quickshear (<i>top</i>); MRI Defacer (<i>bottom</i>) [41]	537
Fig. 21.1	Passive HF RFID tags	552
Fig. 21.2	Passive UHF RFID tags	553
Fig. 21.3	Basic exchange between an RFID tag and reader	559
Fig. 21.4	Using the EPC number to retrieve additional information about the tag and associated asset	561
Fig. 22.1	Block diagram	572

Fig. 22.2	The decision graph leading to ECG segment classification. Given the array y_1, \dots, y_6 , the tree is traversed according to the result of the comparison of the values with 0 in each node, following the true (T) or false (F) edges	574
Fig. 22.3	Garbled circuit scheme	578
Fig. 22.4	Linear selection circuit (part of C) of a node	581
Fig. 22.5	Hybrid LBP protocol. For simplicity we assume that all the y_i values can be packed in a single ciphertext	583
Fig. 22.6	Privacy-preserving ECG diagnosis	584
Fig. 22.7	Classification accuracy of dataset using 21 and 15 features	586
Fig. 22.8	Classification accuracy of dataset using 5 and 4 features	587
Fig. 22.9	A perceptron	591
Fig. 22.10	Transfer functions. (a) <code>tansig</code> . (b) <code>satlin</code>	592
Fig. 22.11	Classification accuracy as a function of the number of nodes in the hidden layer and <code>satlin</code> as activation function	593
Fig. 22.12	Neural network structure. In the ECG classification protocol $n = 4$, $n_h = 6$ and $n_o = 6$	593
Fig. 22.13	Classification accuracy as a function of ℓ^i , ℓ^h , ℓ^o	595
Fig. 22.14	Classification accuracy in function of ℓ^o , with $\ell^i = \ell^h = 13$	595
Fig. 22.15	Hybrid implementation of the neural network	597
Fig. 22.16	Scheme to compute the SNR	600
Fig. 22.17	Sequence of steps performed to evaluate the quality of an ECG signal	604
Fig. 23.1	The participants of healthcare social networks	615
Fig. 24.1	Risks created by the lack of globally harmonisation data privacy standards	645
Fig. 24.2	Three main limitations to anonymisation of personal data	647
Fig. 24.3	Basic principles of national application, Part 2 of the OECD privacy guidelines [56]	649
Fig. 24.4	Basic principles of international application: free flow and legitimate restrictions, Part 4 of the OECD privacy guidelines [56]	650
Fig. 24.5	The UK data protection principles, Part 1 of Schedule 1 of the Data Protection Act 1998 [65]	657
Fig. 24.6	Conditions under which a covered entity is permitted to use and disclose PHI for research purposes [70]	661
Fig. 24.7	The seventeen HIPAA Privacy Rule de-identification fields (U.S. 2014: §164.514(b)(2)(i) [71])	661
Fig. 24.8	HIPAA Privacy Rule limited dataset (U.S. 2014: §164.514(e)(2) [71])	663
Fig. 24.9	Extract from the genome data release and resource sharing policy [31]	669

Fig. 24.10	Wellcome Trust policy on data management and sharing [76]	670
Fig. 25.1	Summary of OCR action on covered entities since 2008.....	684
Fig. 25.2	Implications of Endsley's three-stage model on privacy protection	686
Fig. 25.3	A simplified illustration of information flow in a healthcare setting	689
Fig. 25.4	Example of SA errors in the registration process	690
Fig. 25.5	Level 1 SA error—failure to correctly perceive a situation or misperception of information.....	690
Fig. 25.6	Level 2 SA error—failure to comprehend situation or improper comprehension of information	691
Fig. 25.7	Level 3 SA error—incorrect projection of situation into the future	691
Fig. 26.1	U.S. HIPAA protected health information categories	700
Fig. 26.2	Examples of text de-identification with PHI tagging or resynthesis	701
Fig. 26.3	Workflow supported by MIST and its components	706
Fig. 26.4	The VHA BoB components and text processing pipeline	707
Fig. 26.5	Typical metrics for de-identification applications	711
Fig. 27.1	A fabricated sample medical record before and after surrogate generation	718
Fig. 27.2	Fabricated EMR prior to surrogate generation; PHI are delineated with XML tags	724
Fig. 27.3	Fabricated EMR after surrogate generation; PHI are delineated with XML tags	725
Fig. 27.4	Generic algorithm for replacing alphanumeric strings	726
Fig. 27.5	Algorithm for generating replacement names.....	728
Fig. 28.1	The focus on “identifiability” and “risk” in current data access processes	742
Fig. 28.2	An iterative approach to data access.....	757

List of Tables

Table 2.1	Anonymization algorithms that protect from identity and attribute disclosure (adapted from [20], with permission from Elsevier)	22
Table 2.2	Algorithms that protect from identity disclosure based on demographics (Table adapted from [20], with permission from Elsevier)	26
Table 2.3	Algorithms that protect against identity disclosure based on diagnosis codes (Table adapted from [20], with permission from Elsevier)	27
Table 2.4	Algorithms for preventing attribute disclosure for demographics (Table adapted from [20], with permission from Elsevier)	29
Table 2.5	Algorithms for preventing attribute disclosure based on diagnosis codes (Table adapted from [20], with permission from Elsevier)	30
Table 3.1	A sample set-valued dataset [5]	49
Table 4.1	Original prescription database (derived from [11])	63
Table 4.2	Disclosed prescription database ($k = 2$) (derived from [11])	63
Table 4.3	Odds ratios (OR) and statistical significance (SS) for males (M) and females (F) in original data set (kisbq18).....	70
Table 4.4	Odds ratios (OR) and statistical significance (SS) for males (M) and females (F) in original data set (kisbq20).....	71
Table 4.5	Logistic regressions for each dataset	76
Table 5.1	An example of an RT-dataset containing patient demographics and diseases	84
Table 5.2	(a) A 2-anonymous dataset with respect to relational attributes, (b) a 2^2 -anonymous dataset with respect to the transaction attribute, and (c) a $(2, 2^2)$ -anonymous dataset...	85
Table 5.3	Comparison of data anonymization tools	87

Table 5.4	A $(2, 2^2)$ -anonymous dataset with privacy constraints $\mathcal{P} = \{\text{Flu}, \text{Herpes}\}$ and utility constraints $\mathcal{U} = \{\{\text{Asthma}, \text{Flu}\}, \{\text{Herpes}, \text{Eczema}\}\}$	95
Table 5.5	Overview of the methods supported from SECRETA	101
Table 5.6	(a) A fictitious medical dataset presenting the Year-Of-Birth, Sex, Race, and Diagnosis codes of patients, (b) mapping of attribute Sex, and (c) mapping of attribute Race	102
Table 6.1	Example dataset and the result of applying the transformation (1,0)	117
Table 6.2	Datasets used in the experiments	140
Table 6.3	Runtime measures for risk-based anonymization	140
Table 6.4	Runtime measures for 5-anonymity	140
Table 6.5	Utility measures for risk-based anonymization	141
Table 6.6	Utility measures for 5-anonymity	141
Table 7.1	(a) Dataset comprised of diagnosis codes, and (b) diagnosis codes contained in the dataset of Table 7.1a and their description (reprinted from [15], with permission from Elsevier)	151
Table 7.2	(a) Utility requirements, and (b) anonymized dataset that <i>does not</i> satisfy the utility requirements (reprinted from [15], with permission from Elsevier)	153
Table 7.3	Anonymized counterpart of the dataset in Table 7.1a, using the utility-constrained approach (reprinted from [15], with permission from Elsevier)	154
Table 7.4	Mappings between diagnosis codes and generalized terms, created by set-based generalization	157
Table 7.5	A possible dataset reconstructed from the dataset of Table 7.3 (reprinted from [15], with permission from Elsevier)	159
Table 7.6	Examples of hierarchy-based utility constraints	163
Table 7.7	Disassociation with a shared chunk (reprinted from [15], with permission from Elsevier)	169
Table 7.8	Sets of diagnosis codes that are added into the priority queue of CBA, and their support	170
Table 7.9	Anonymized dataset by CBA using the utility policy of Table 7.2a	171
Table 7.10	Disassociated dataset with a shared chunk (reprinted from [15], with permission from Elsevier)	172
Table 7.11	The result of functions M_O and M_A , for CBA and for a reconstructed dataset, produced by DIS	172
Table 7.12	MRE scores for each utility constraint in Table 7.2a	173

Table 7.13	Average percentage of records that are retrieved incorrectly, for workloads having different δ values and for: (a) \mathcal{U}_1 , (b) \mathcal{U}_2 , and (c) \mathcal{U}_3	174
Table 8.1	An illustrative example of a composition attack; the shared common record is revealed by intersecting two corresponding equivalence classes	181
Table 8.2	An illustration of differential privacy based publications of datasets in Table 8.1. Anemia, Cancer, Migraine, Diabetes and Cough are all sensitive values in the datasets. The counts of sensitive values are noised and published with the equivalence class. It is difficult for an adversary to find true common sensitive values using noised counts. Note that the counts are small since we use the same datasets from Table 8.1	186
Table 8.3	Comparison of risk of composition attack of two equivalence classes: Case 1	188
Table 8.4	Comparison of risk of composition attack of two equivalence classes: Case 2	189
Table 8.5	Domain size of different attributes	190
Table 9.1	Disclosure risk and information loss for the generated table	225
Table 10.1	Overview of the PPRL techniques reviewed in the chapter.	240
Table 10.2	Example of k -anonymity, with $k = 2$	242
Table 10.3	The Soundex encoding	250
Table 13.1	The responsibilities of the CEI centers	317
Table 13.2	The profile of the study cases and the selected sample to build the gold-standard (reprinted from [52], with permission from Elsevier)	338
Table 13.3	Assignment of sample cases to judges and results from the first round review ^a	342
Table 13.4	Comparison between the enhanced RBAC model and the CEIAdmin system	345
Table 13.5	Measuring the effectiveness of the enhanced RBAC model and the CEIAdmin system with a gold-standard	346
Table 15.1	Overview of the approaches employed to maintain confidentiality	403
Table 15.2	Overview of the approaches employed to maintain data integrity	405
Table 15.3	Overview of the approaches employed to offer collusion resistance	408
Table 15.4	Overview of the approaches employed to maintain anonymity ..	411

Table 15.5	Overview of the approaches employed to maintain authenticity	413
Table 15.6	Overview of the approaches employed to maintain unlinkability	414
Table 16.1	Key aspects of GWAS and genomic data sharing (GDS) policies	429
Table 17.1	Raw genome data	447
Table 17.2	Genome data partitioned into blocks	450
Table 17.3	Anonymous data ($\epsilon = 1, h = 2$)	452
Table 17.4	Data utility of chr2 data set with privacy budget of 1.0 and power of 0.01	455
Table 17.5	Data utility of chr10 data set with privacy budget of 1.0 and power of 0.09	455
Table 18.1	Frequently used notations	467
Table 20.1	Landmarks used in facial reconstruction [13, 15, 36, 52] and corresponding feature memberships.....	539
Table 21.1	The frequency band, RFID type, corresponding ISO standard, and typical applications	553
Table 22.1	Classification pattern (“*” means that the value of the variable does not influence the classification)	573
Table 22.2	QDF-based classification results (results from [1, Chap. 8])	575
Table 22.3	QDF-based classification results obtained in the tests	575
Table 22.4	Protocols for secure evaluation of private LBPs	582
Table 22.5	Estimated communication complexity of LBP with QDF	588
Table 22.6	Performance of protocols for secure ECG classification through LBP	589
Table 22.7	Inputs to the GC implementation of the neural network	596
Table 22.8	Complexity of NN-based ECG classification protocol	597
Table 22.9	Number of bits necessary to represent the values obtained by a worst case analysis	602
Table 22.10	SNR protocol data transfer	603
Table 22.11	Maximum value and number of bits necessary for the magnitude representation of the variables involved in the computation by worst case analysis	607
Table 22.12	Bandwidth (bits) required by the protocol	607
Table 22.13	Performance of the protocol using the linear classifier or a single feature	608
Table 23.1	Identification techniques that may affect medical privacy (based on [10])	619
Table 23.2	Privacy principles and their application is HSNs.....	623

Table 23.3	A list of privacy threats and associated risks in HSNs	623
Table 24.1	HIPAA Safe Harbor example de-identification of a simple medical dataset prior to de-identification and following de-identification (changed/removed data in bold)	662
Table 24.2	Deadlines for data submission and release in a supplement to the NIH Genomic Data Sharing Policy, which apply to all large-scale, NIH-funded genomics research	667
Table 25.1	The 12 largest privacy breaches identified by the OCR.....	683
Table 26.1	Advantages and disadvantages of methods used for text de-identification	702
Table 26.2	Sensitivity of machine learning-based text de-identification applications (partly reproduced from [19])	711
Table 26.3	Overall (micro-averaged) performance at the PHI level (partly reproduced from [19])	712
Table 26.4	Sensitivity of de-identification applications when generalized to a different type of clinical notes	712
Table 27.1	HIPAA's list of PHI categories	723
Table 28.1	A comprehensive framework for proportionate governance: domains for adjunction and associated determination of risk	753
Table 28.2	The adjudication scenario 1	758
Table 28.3	The adjudication scenario 2	759

Chapter 1

Introduction to Medical Data Privacy

Aris Gkoulalas-Divanis and Grigoris Loukides

Abstract The advancements in medical and information technology have resulted in a tremendous increase in the amount and complexity of medical data that are being collected. These data are a valuable source for analyses that have strong potential to improve both medical research and practice. However, such analyses have also raised considerable concerns over potential violations of privacy and misuse of medical data. To address such concerns, technological and procedural solutions are necessary. These solutions must be applicable to different types of data, ranging from patient demographics to medical images, and be able to meet diverse application requirements, such as data publishing and health information exchange. This chapter provides an introduction to the field of medical data privacy, offers a taxonomy of the different research directions, and presents an overview of the state-of-the-art privacy-preserving solutions.

1.1 Introduction

Recent developments in medical and information technology allow the collection of massive amounts of medical data. These data contain demographic, genomic, and clinical information, including diagnoses, medications, medical images and laboratory test results. For instance, electronic health records are utilized by multiple parties with different roles and responsibilities, as a result of the increasing adoption of Electronic Medical Record/Electronic Health Record (EMR/EHR) systems [6, 15]. The use of the collected data is a valuable source for analyses that significantly benefit both medical research and practice, leading to effective ways of preventing and managing illnesses, as well as the discovery of new drugs and therapies. For example, it is necessary to share large amounts of demographic, clinical, and genomic data among collaborating healthcare institutions, in the

A. Gkoulalas-Divanis (✉)

Smarter Cities Technology Centre, IBM Research, Dublin, Ireland
e-mail: arisdiva@ie.ibm.com

G. Loukides

School of Computer Science & Informatics, Cardiff University, Cardiff, UK
e-mail: g.loukides@cs.cf.ac.uk

context of genomic studies [16], which are considered as an important step towards the realization of personalized medicine. The quality of healthcare that patients receive can also be improved as a result of the increased availability of medical data to clinicians [15].

At the same time, however, there are considerable concerns regarding potential violations of privacy and misuse of medical data. For example, according to a recent survey [19], 59 % of patients believe that the widespread adoption of EMR/EHR systems will lead to more personal information being lost or stolen, while 51 % believe that the privacy of health information is not currently sufficiently protected. In addition, various public campaigns have urged people to opt-out and to dissent to the use of their medical records [5, 18], and as a result large efforts for improving health services, such as *care.data* in England, have been significantly delayed [21].

To address these concerns, it is important to consider the entire lifecycle of medical data when protecting them. This implies that medical data must be collected and communicated securely, be accessed only by authorized parties, and not disclosing any private and/or sensitive medical information when disseminated. To meet these objectives, a large number of technological solutions have been proposed. These include formal privacy models, algorithms, protocols, languages, architectures, and software tools. In addition, privacy protection is a legal requirement, which is posed by worldwide legal frameworks, including the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) [11] in the United States, and the Data Protection Act [28] in the United Kingdom. Closely related to the enforcement of these policies, are recent medical data governance efforts. These include the establishment of the so-called *safe havens*, i.e., environments which allow researchers to securely access medical data. Examples of safe havens are the *Population Data BC* [23] in Canada and the *Farr Institute* [8] in the United Kingdom.

These efforts have shaped the area of medical data privacy. Broadly speaking, the research in this area falls in one of the categories described in the following subsections.

1.1.1 Privacy in Data Sharing

Protecting the privacy of data prior to their sharing is a central topic in medical data privacy. Privacy-protected data may either become accessible by authorized parties or publicly available. For example, medical data are often accessed by researchers, who perform various types of analyses, spanning from simple statistical tests to advanced data analytics, in the context of their studies.

The analysis of data in a privacy-preserving way is possible through *statistical disclosure control* methods, which modify aggregated data prior to analysis, or through *output perturbation* methods, which modify the analysis result. In addition, medical data that have been collected as part of funded research projects need to be deposited into central repositories, in order to enable the validation of

existing results as well as to be used in novel future studies. In this setting, data *anonymization* methods that transform the original medical data to satisfy privacy, while preserving data utility, are necessary. The variety of medical data types and analytic tasks that are applied to the data call for different approaches. As an example, anonymization methods for demographic data are not applicable to data containing patients' diagnosis codes, since these two types of data are modeled and analyzed in substantially different ways. Similarly, the anonymization of patients' longitudinal data and the de-identification of clinical reports call for vastly different solutions.

1.1.2 Privacy in Distributed and Dynamic Settings

Numerous applications require the protection of data that are distributed into multiple parties. These parties hold, and aim to integrate, parts of data about the same patient, or about different patient subpopulations. In this case, *record linkage* methods are necessary to securely construct a “global” view of the data. These methodologies often employ secure multiparty computation (SMC) protocols to construct this “global” view, without the sharing of original (unencrypted) data among parties. The goal of these methods is to achieve data integration in a way that no party, who participates to the protocol, learns the data that is held by any other party (as well as any other sensitive information).

In addition, patients' medical information often needs to be exchanged between different parties, in a privacy-preserving way. This offers multiple benefits, reducing the possibility of medical identity theft, financial fraud, and errors. To achieve this, *access control* methods, which manage how these parties access information are necessary. These methods determine when and how a subset of data about a patient can be accessed by specific parties, which perform a certain task or have a certain role. An example of an effective approach to managing the access to data is based on the notion of *consent*. This approach empowers patients with the capability of granting and revoking access to their (medical) data, in a way that changes over time and is specified according to the context. Furthermore, medical data may be stored or processed in a *cloud* infrastructure. This poses various privacy threats, such as malicious data access and intentional data modification, which have to be appropriately addressed.

1.1.3 Privacy for Emerging Applications

Advances in sequencing technology, telemedicine, signal processing, and social networks allow the collection of large amounts of complex data. Examples are DNA sequences, medical images, biomedical signals, and health social network data. All these types of data contain confidential or sensitive information about

patients and need to be privacy-protected. This is far from straightforward, due to the inherent complexity of these data types. For instance, if the genomic sequence of a patient is not sufficiently protected, privacy-intrusive associations between the patient and family members may be disclosed. Another significant challenge involved in the protection of these data is efficiency. This is true for methods that employ cryptographic operations, which are not generally scalable to large datasets.

To protect genomic data, techniques such as *noise addition*, which is used to enforce an appropriate privacy principle, or *cryptography*, which is used to securely compute a result on the data, are typically employed. The goal of both of these techniques is to maintain a desired level of privacy protection, while preserving the utility of data in supporting genomic studies.

In addition, the sharing of medical images is promoted by various initiatives worldwide, such as the Picture Archiving and Communications Systems (PACS) in the United Kingdom [22], and the Canada Health Infoway-Inforoute [4]. The protection of medical image data can be achieved by cryptographic means (mainly *encryption* and *digital signatures*), while techniques such as *watermarking* protect the images from unauthorized modification and copying.

Another class of emerging applications for both disease diagnosis and treatment is based on biomedical signal processing. An important requirement in these applications is to be able to process a biomedical signal without revealing its content or the processing results, to unauthorized users. This is achievable by performing analysis directly on the encrypted signal.

Furthermore, health social networks allow their users (e.g., doctors, patients and carers) to communicate and interact, in order to virtually manage and improve the health of patients. Health social networks and related communities can either be built on top of existing social networking applications or on dedicated platforms. Protecting the data that are deposited by patients, or result from the communication and interaction among users, is particularly important. This is possible by building solutions using a multitude of technologies that were mentioned before, including *anonymization*, *cryptography*, and *access control*.

1.1.4 Privacy Through Policy, Data De-identification, and Data Governance

Clinical information systems, such as web-based electronic healthcare systems, systems for personal health monitoring, and physician order entry systems, are increasingly adopted and have already simplified and improved many aspects of healthcare.

These systems contain sensitive information that must be managed in a privacy-preserving way. To achieve this, adherence to legal frameworks, such as HIPAA and the Data Protection Act, is an important and mandatory step. These frameworks specify the responsibilities of organizations regarding the privacy protection of

personal health information. However, complying with these frameworks is both challenging and costly for a healthcare organization. Furthermore, discovering individually identifiable health information about patients is not straightforward, especially when data are unstructured (i.e., in a textual form). Thus, techniques which *de-identify* text data (i.e., detect, and remove or modify personal identifying information) are necessary. In addition, clear data governance processes are needed to deal with the privacy-preserving management of clinical information.

This handbook covers the above-mentioned research areas of medical data privacy in a comprehensive way. Each area corresponds to a different part of the book.

In the remaining of this chapter, we discuss each of these areas in more detail and provide an overview of the contents of the book. Specifically, Sects. 1.2 and 1.3 discuss privacy protection in the context of data sharing, and in distributed and dynamic settings, respectively. Section 1.4 presents a range of emerging privacy-preserving applications. Section 1.5 focuses on the issues of policy, data de-identification, and data governance, which are central to the management of medical information. Last, Sect. 1.6 concludes the chapter.

1.2 Part I: Privacy in Data Sharing

This part of the book focuses on preserving privacy in data sharing. Specifically, it explains how this can be achieved through data anonymization techniques. These techniques are based on various formal privacy principles and data transformation operations, and they are applied by a series of methods and software tools. The part concludes with a survey of statistical disclosure control methods for medical data.

Anonymization algorithms are very important, in the context of data sharing, and their development has attracted significant attention. As a result, a large number of anonymization algorithms for medical data have been proposed recently. Chapter 2 reviews the most popular of these algorithms and explains their objectives, as well as the main aspects of their operation. Furthermore, it presents several directions for future research in the area of developing data anonymization algorithms. Interested readers may also refer to [10] for a more comprehensive coverage of this area.

Differential privacy [7] has emerged as a strong privacy principle that can be used to protect medical data, prior to their sharing. Intuitively, this privacy principle ensures that any inferences that can be made by an attacker do not depend significantly on the presence or absence of information about a patient (e.g., a patient's record). In the recent years, many approaches for enforcing differential privacy on medical data have been proposed. Chapter 3 introduces differential privacy and provides an extensive survey of these approaches with focus on private histogram generation and synthetic data publication.

An important consideration when applying differential privacy in practice is the preservation of data utility. In fact, the enforcement of differential privacy is typically performed through (Laplace) noise addition, which may significantly

degrade data utility. Several works have indicated this limitation of differential privacy when compared to syntactic anonymization methods. At the same time, research has shown that in certain data publishing scenarios, differential privacy can achieve reasonable levels of data utility. Chapter 4 studies the issue of assessing and maintaining the utility of differential privacy, in the context of a fundamental statistical task (logistic regression [9]). The analysis and the experimental results presented in this work confirm that the production of differentially private data with high utility is challenging, particularly when the data are high-dimensional.

Another important privacy principle to protect medical data is k -anonymity [25, 26]. k -anonymity belongs to the family of *syntactic* privacy-preserving data publishing approaches and is probably the first formal mathematical model that has been proposed for offering anonymity. The k -anonymity principle aims to prevent an attacker from associating a patient with their patient record in a published dataset. This attack is referred to as *identity disclosure* and constitutes a major privacy breach. Identity disclosure is possible when patient data (e.g., demographic values or diagnosis codes) are contained in both the published dataset and in publicly/widely available external data sources (e.g., voter registration lists, Census data, and hospital discharge summaries). A triangulation of the patient data with those external data sources may lead to unique patient records for some of the patients whose information is recorded in the dataset, hence reveal their identity. To prevent identity disclosure, k -anonymity requires modifying the data (usually through selective data generalization and/or suppression), prior to their sharing, so that each record in the published dataset can be associated with at least k different individuals, where k is a parameter specified by data owners. This limits the probability of performing identity disclosure to at most $\frac{1}{k}$. We note that in medical research, typical values of k are $k = 3$ and $k = 5$, which are considered to offer a sufficient level of anonymity.

Various algorithms for enforcing k -anonymity on medical data have been proposed [10]. Recently, several of these algorithms have been implemented into software systems. This is important because the implementation and application of anonymization methods requires significant expertise. Chapter 5 presents an overview of a system called *SECRETA*, which was developed by the University of Peloponnese in collaboration with Cardiff University and IBM Research. The system allows evaluating and comparing nine popular data anonymization algorithms that are applicable to patient data containing demographics and/or diagnosis codes. The application of the algorithms is performed, in an interactive and progressive way, and the results, including attribute statistics and various data utility indicators, are summarized and presented graphically. Another anonymization system, called ARX, was developed by the Institute for Medical Statistics and Epidemiology of the University of Technology Munich (TUM). Chapter 6 discusses this anonymization system, which is applicable to patients' demographics. The system supports various algorithms based on k -anonymity, as well as on other popular principles [12, 20] that enhance the protection provided by k -anonymity. An important feature of ARX is that it guides the user, through all steps of the anonymization process, using several helpful wizards.

k -Anonymity offers a good solution to data privacy, by providing formal privacy guarantees. Applying, however, k -anonymity on high dimensional data, such as patients' diagnosis codes, is challenging. In response, different principles that are based on k -anonymity have been proposed [13, 24]. The main idea behind these privacy principles is to prevent the use of specific combinations of diagnosis codes in identity disclosure attacks. The specification of these combinations can be performed by data owners, or automatically [13]. Thus, the application of these principles avoids overprotecting the data and unnecessarily reducing data utility. Chapter 7 discusses the application of these principles through two popular operations, namely data *generalization* [13, 25, 26] and *datadisassociation* [14, 27]. Generalization replaces diagnosis codes with sets of diagnosis codes (e.g., diabetes type I may be replaced by a set containing diabetes-I and diabetes-II). On the other hand, *disassociation* partitions the diagnosis codes into groups to conceal their associations. The chapter also reviews state-of-the-art algorithms based on generalization and on disassociation.

Most of the syntactic approaches that have been proposed to facilitate privacy-preserving data publishing assume a single data release, where the original dataset is first anonymized and then made available to the intended recipients. After its original release, the data is no longer updated with (anonymized) information about new individuals or additional information about the existing ones. Moreover, a subset of the same data is typically not used as part of other anonymous data releases, as this could lead to privacy disclosures. Chapter 8 is devoted to this important issue of preventing *composition attacks* in the privacy-preserving sharing of medical data. These attacks involve the intersection of independently published datasets, to infer sensitive information (e.g., that a patient is diagnosed with a mental disorder). To mitigate these attacks, this chapter presents two different approaches that are applicable to patient demographics. The first approach is based on noise addition and the second on data generalization. The chapter also presents analysis and experiments to demonstrate the strengths and weaknesses of each approach.

The approaches discussed so far focus on medical data that are being shared by healthcare organizations. However, statistical agencies also share medical data that are collected by individuals in surveys, censuses, and registers. These data often contain sensitive information (e.g., statistics on abortions or diseases) and must be protected prior to their dissemination. There are, however, important challenges in doing so, which stem from the complexity of data (e.g., when data are longitudinal) or the requirement for publicly releasing data as part of government open data initiatives. Chapter 9 surveys different approaches that are employed by statistical agencies to protect the privacy of health data. These approaches are either applicable to individual-specific data, or to aggregate data. In addition, the chapter discusses approaches that are currently being assessed by statistical agencies.

1.3 Part II: Privacy in Distributed and Dynamic Settings

The second part of the book discusses privacy protection in distributed and dynamic settings. It first surveys techniques for privacy-preserving record linkage, as well as their implementation and use by dedicated centers for linking medical data that have been recently established around the world. Subsequently, the important issue of privacy preservation in Health Information Exchange (HIE) systems is discussed. After surveying threats and solutions in these systems, techniques for access control and consent management are presented in detail. The part concludes with a survey of privacy issues related to the management of medical data in cloud environments.

Many large-scale medical research studies require data about the same patient, which are distributed into (i.e., held or managed by) different parties. For example, social data about patients are typically collected by statistical agencies, while the patient's clinical data are collected by one or more hospitals that the patient has visited. To enable these studies, it is important to construct a holistic view of the patients' health information, by combining all the available data referring to the same patient, from the different data sources. In addition, the construction of this (holistic) record for each patient should be performed without disclosing the data that is held by each party, due to privacy concerns (i.e., the parties cannot simply share all their data). To achieve this, privacy-preserving record linkage methods have been proposed. Chapter 10 provides a comprehensive survey of privacy-preserving record linkage methods, with a focus on those that use identifying information (e.g., names) to perform the linkage. The latter methods preserve individual patients' privacy and are efficient and effective.

Privacy-preserving record linkage is of great practical importance. This is evidenced by the establishment of dedicated medical record linkage centers, in various countries, including the United Kingdom, Canada, and Australia [3]. These centers routinely collect and link medical data, as well as provide secure access to the data to trusted researchers. Chapter 11 discusses both procedural (i.e., practices and processes) and technical (i.e., architectures and protocols) approaches that are typically employed by record linkage centers.

Another important setting in which data must be privacy-protected involves Health Information Exchange (HIE) systems accessed by multiple parties. These systems contain patient information that is directly collected from patients (e.g., from physicians or pharmacies). HIE systems are highly complex and dynamic. For example, they may be federated to allow broad accessibility that spans across multiple healthcare providers. HIE systems improve quality of care and patient safety, and they significantly reduce medical errors. However, the increased availability of patient information to multiple parties poses serious threats, such as identity theft and medical financial fraud. Chapter 12 surveys issues related to privacy in the context of HIE systems, and it presents principles that users and administrators of these systems should adopt in order to preserve patient privacy.

The information contained in HIE systems needs to be accessible by multiple collaborating parties, for purposes ranging from team-based patient care to clinical

education. The main challenge in controlling the access to this information is the dynamic nature of HIE systems. Specifically, there are changes to the data, to the level of data access, and to the specific (and often conflicting) requirements of the different parties. Chapter 13 surveys the state-of-the-art research on access control to support team collaboration and workflow management. In addition, it introduces the New York State HIV Clinical Education Initiative (CEI), an application that requires complex information access in the combined contexts of workflow and team collaboration.

In addition, information access in HIE systems needs to be performed in accordance with patients' consent [17, 30]. This calls for *consent management* approaches, which aim to capture the specific requirements of patients and health-care providers for access to patients' medical data, in a way that reduces the risk of errors and preserves privacy. Consent management is a dynamic and context-specific process, since patients may want to handle consent in accordance with the actual situation. Chapter 14 presents an approach for managing consent that is based on *authorization policies* (i.e., rules that govern the access to a patient's data). The approach is able to simplify the management of consent, by combining and collectively managing policies that have a common aim, in a specific context.

Recently, there is tremendous interest for utilizing cloud computing infrastructures to store, manage, analyze and share medical data. This offers various benefits, including scalability, increased availability of information, and cost effectiveness. However, outsourcing sensitive medical data to third-party cloud providers may raise serious privacy concerns. Chapter 15 provides an overview of the privacy issues related to the storage and sharing of medical data in the cloud. Subsequently, it focuses on effective strategies to mitigate these concerns. Each of these strategies is evaluated and open issues for future research are presented. We note that this is a very interesting and largely unexplored research area, which asks for novel methods that can offer sufficient privacy protection with guarantees for the stored medical data, in order to increase the adoption of cloud computing on healthcare.

1.4 Part III: Privacy for Emerging Applications

In this part of the book, privacy protection in a range of emerging medical applications is discussed. This includes applications featuring the management, sharing, and analysis of: (a) genomic data, (b) medical images, (c) Radio Frequency IDentification (RFID) data, (d) biomedical signals, and (e) health social network data. Due to the differences in their semantics and use, these types of data are susceptible to different privacy attacks and require very different approaches to be protected.

Genomic data are highly important for supporting research in personalized medicine, as well as for improving the accuracy of diagnosis and public health. However, they are probably the most sensitive type of medical data as they may reveal a patient's sensitive information, including identity, ethnic group, and disease

association, as well as information about the ancestors and descendants of the patient. Thus, the sharing and analysis of patient genomic data must be performed in a way that provably preserves privacy. Chapter 16 introduces the state-of-the-art in genome privacy research. After providing an overview of the legal, ethical and technical aspects of genome privacy, the authors present various attacks and solutions that have been proposed for preserving the privacy of genomic data, as well as discuss several directions for future research in this hot research area.

Enforcing differential privacy to protect the privacy of patients' genomic data is an important approach, particularly because differential privacy makes few assumptions about the background knowledge of attackers. This helps to eliminate inferences that are difficult to predict, but can have serious consequences to privacy. Chapter 17 presents an effective algorithm that follows this approach. The algorithm splits raw genome sequences into blocks, subdivides the blocks in a top-down fashion, and finally adds noise to counts, in order to enforce differential privacy. Importantly, the algorithm is able to preserve data utility in supporting different genome data analysis tasks, as demonstrated by the authors through experiments.

The second major approach to protecting the privacy of genomic data is based on *cryptography*. This approach assumes that a set of authorized users are interested in obtaining the results of a medical analytic task, instead of obtaining the entire dataset. Chapter 18 discusses this approach and adopts it to address four interesting problems. These are (a) the management of raw genomic data, (b) the use of genomic data for conducting medical tests, (c) the use of genomic data for genetic association studies, and (d) the secure data storage and retrieval of genomic data. The problems are addressed using methods based on cryptographic primitives, such as *order-preserving* [1] and *homomorphic* [2] encryption.

Medical images are very important in diagnostic applications, and they need to be protected from a range of attacks, including unauthorized view, modification, and copying, to ensure patient confidentiality. Chapter 19 surveys approaches that combine *encryption* with *watermarking* to achieve this. Encryption guarantees that medical images cannot be viewed by unauthorized parties, while watermarking prevents the copying of images from authorized parties that have decrypted them. The chapter presents a range of methods that combine encryption and watermarking, and then focuses on a practical algorithm for joint encryption-watermarking.

In addition, guarding against patient re-identification (i.e., disclosing the identity of patients, from an image that is devoid of directly identifying information) is necessary from both a legal and an ethical point of view. In the context of image data, patient re-identification occurs when an attacker *verifies the identity of the owner* of the image, or when the attacker *creates an accurate profile for the potential owner* of the image. Chapter 20 discusses the threat of re-identification in the context of neuroimage data. After presenting data properties and processes that are useful for protecting neuroimages, the authors discuss privacy risks relevant to neuroimage data. Subsequently, the chapter presents an effective solution for neuroimage data that is able to satisfy legal and regulatory requirements.

Next, we consider Radio Frequency IDentification (RFID) data. These data are increasingly used to improve supply-chain management and service offerings, by

monitoring the locations of patients and healthcare providers. The data that is stored in RFID devices may contain sensitive information about patients that can be obtained, for example, by intercepting the transmission used to track a patient. In addition to obtaining access to sensitive patient data, there are more privacy attacks that need to be considered for RFIDs, such as data modification, interruption, and fabrication. Chapter 21 surveys these attacks, along with further issues related to privacy, in the context of various real applications of RFIDs in healthcare. Furthermore, it discusses a range of potential solutions to each of these threats.

The diagnosis of various heart-related disorders, including ventricular contractions, fibrillation and tachycardia, is based on the analysis (classification) of a type of biomedical signal, referred to as ElectroCardioGram (ECG). Chapter 22 presents three algorithms for classifying encrypted ECG signals. This is a challenging task, which requires preserving the representation of the signal after its encryption, as well as the intermediate computation results. The privacy algorithms presented in this chapter are based on different primitives (classification functions) and are evaluated using real data, in terms of their achieved accuracy and efficiency.

Another class of emerging applications in medicine are based on health social networks. In these applications, users (i.e., doctors, patients, caregivers, etc.) exchange data of various forms, such as text and images, to provide or receive emotional support, share advice and experiences, and find medical information that would help them address a medical condition. Thus, the use of health social networks can help towards improving patients' health and quality of life. On the other hand, however, health social network data are susceptible to privacy attacks, such as re-identification, discrimination, profiling, and sensitive information disclosure. Chapter 23 provides a comprehensive discussion of these attacks, in the context of health social networks of different types, as well as the implications of these attacks to patients. Furthermore, it presents tools and methods that can be used to guard against these attacks.

1.5 Part IV: Privacy Through Policy, Data De-identification, and Data Governance

Part IV of the book reviews important issues related to privacy-preserving data management in healthcare systems. The first issue regards compliance with existing legal frameworks. This is a legal requirement for healthcare institutions and organizations (at large) that hold electronic health records, but it is far from straightforward to achieve. One major challenge is to effectively *de-identify* patients' medical data. On this end, the US legislation (HIPAA) provides a set of guidelines for the removal of Protected Health Information (PHI) from the data, where PHI is any information in a medical record that can be used to identify an individual. This challenge becomes more important in the context of unstructured medical data (text), such as doctors' clinical notes and prescriptions, where PHI have to be first discovered in the data

and then redacted. Another major challenge is that the existing legal frameworks are generic and sometimes abstract. Consequently, *data governance* strategies are needed to support compliance to these frameworks. Both of the aforementioned challenges are discussed extensively in this part of the book, and methods for addressing them are presented.

Legal frameworks that have been proposed by the governments of many countries (e.g., United States, United Kingdom, and Canada), as well as by major research funding organizations (e.g., the NIH, Wellcome Trust, and Genome Canada) are surveyed in Chap. 24. These frameworks are comprised of laws, policies, and regulations, and are an important step for preserving data privacy. The content, significance, and practical usefulness of these frameworks is explained, and barriers to facilitating global data sharing are identified. The chapter also suggests ways to overcome these barriers, while maintaining robust data-privacy protection.

The breach of privacy laws, policies, and regulations is often caused by human errors and incurs significant costs to healthcare organizations. Chapter 25 categorizes the human errors and explains how they may lead to privacy breaches, in the context of the HIPAA privacy policy [11], a key legal framework for medical data in the United States. In addition, it provides an analysis of various cases of privacy breaches and reports on the most important cases of HIPAA privacy breaches in which the U.S. Office for Civil Rights has reached a resolution agreement. The analysis identifies the latent causes of privacy breaches and has important implications in policy formulation and system design.

As mentioned earlier, the de-identification of medical text is necessary to comply with legal policies and regulations. In addition, de-identification is essential to facilitate the secondary use of clinical information, because a significant amount of clinical information is contained in text. While manual de-identification is possible, it is a tedious, costly, and error-prone process. This led to the development of automated de-identification approaches, which are more principled and efficient. Chapter 26 introduces the problem of medical text de-identification, and then surveys and evaluates existing methods, with a focus on those that have been customized for the U.S. Veterans Health Administration [29].

A class of text de-identification methods replace PHI with terms that can be read naturally, which are called *surrogates*. For example, instead of replacing the doctor's name, "Dr. Riley" with a pseudonym "Doctor1" (or redacting it), they replace it with another plausible name, e.g., "Dr. Robertson". This is important for de-identified data intended to be read naturally, such as hospital discharge summaries and correspondence between doctors. Chapter 27 presents a de-identification method based on the generation of surrogates. It also discusses the requirement of preserving important relationships between terms (e.g., co-reference) in the data, which is needed in order to generate realistic surrogates.

The second major challenge to comply with policies and regulations is data governance. This is a procedural requirement, which, in the context of medical data, aims to help a group of healthcare organizations to achieve a collective goal, such as performing a research study. Data governance becomes increasingly important in cases when health data are collected by different parties (e.g., healthcare institutions,

clinics, and pharmacies) and need to be integrated with other forms of data (e.g., social data). Chapter 28 provides a comprehensive discussion of medical data governance. This includes current practice and associated challenges. In addition, it presents a framework for medical data governance, which aims at enhancing privacy while supporting medical data analysis.

1.6 Conclusion

Medical data privacy is a research area with a broad spectrum of applications, ranging from genomics to patient monitoring. This area has attracted significant interest from the computer science, medical informatics, and statistics communities. This chapter provided an overview of the major topics in medical privacy. The first three topics cover aspects of privacy preservation in data sharing, distributed and dynamic settings, and emerging applications. These include anonymization, statistical disclosure control, record linkage, access control, and secure computation. The fourth topic focuses on privacy aspects related to the management of patient information. These include legal frameworks for privacy preservation, de-identification of medical text, and data governance.

References

1. Agrawal, R., Kiernan, J., Srikant, R., Xu, Y.: Order preserving encryption for numeric data. In: SIGMOD, pp. 563–574 (2004)
2. Ateniese, G., Fu, K., Green, M., Hohenberger, S.: Improved proxy re-encryption schemes with applications to secure distributed storage. ACM Trans. Inf. Syst. Secur. **9**(1), 1–30 (2006)
3. Boyd, J.H., Ferrante, A.M., O’Keefe, C.M., Bass, A.J., Randall, S.M., Semmens, J.B.: Data linkage infrastructure for cross-jurisdictional health-related research in Australia. BMC Health Serv. Res. **12**(480), 10pp. (2012)
4. Canada health infoway-inforoute. <http://www.infoway-inforoute.ca> (2015). Accessed 6 Sept 2015
5. Care.data. <http://www.care-data.info/> (2015). Accessed 6 Sept 2015
6. Dean, B., Lam, J., Natoli, J., Butler, Q., Aguilar, D., Nordyke, R.: Use of electronic medical records for health outcomes research: A literature review. Med. Care Res. **66**(6), 611–638 (2010)
7. Dwork, C.: Differential privacy. In: ICALP, pp. 1–12 (2006)
8. Farr Institute. <http://www.farrinstitute.org/> (2015). Accessed 6 Sept 2015
9. Freedman, D.A.: Statistical Models: Theory and Practice. Cambridge University Press, Cambridge (2009)
10. Gkoulalas-Divanis, A., Loukides, G., Sun, J.: Publishing data from electronic health records while preserving privacy: a survey of algorithms. J. Biomed. Inform. **50**, 4–19 (2014)
11. HIPAA privacy rule. <http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/> (2015). Accessed 6 Sept 2015
12. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: ICDE, pp. 106–115 (2007)

13. Loukides, G., Gkoulalas-Divanis, A., Malin, B.: Anonymization of electronic medical records for validating genome-wide association studies. *Proc. Natl. Acad. Sci.* **107**(17), 7898–7903 (2010)
14. Loukides, G., Liagouris, J., Gkoulalas-Divanis, A., Terrovitis, M.: Disassociation for electronic health record privacy. *J. Biomed. Inform.* **50**, 46–61 (2014)
15. Makoul, G., Curry, R.H., Tang, P.C.: The use of electronic medical records communication patterns in outpatient encounters. *J. Am. Med. Inform. Assoc.* **8**(6), 610–615 (2001)
16. Manolio, T., Collins, F., Cox, N., Goldstein, D., Hindorff, L., Hunter, D.: Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009)
17. Marinovic, S., Dulay, N., Sloman, M.: Rumpole: an introspective break-glass access control language. *ACM Trans. Inf. Syst. Secur.* **17**(1), 1–32 (2014)
18. medConfidential: keep my secrets <https://medconfidential.org> (2015). Accessed 6 Sept 2015
19. National partnership for women & families, making it meaningful: how consumers value and trust health it survey. <http://www.nationalpartnership.org/> (2015). Accessed 6 Sept 2015
20. Nergiz, M.E., Atzori, M., Clifton, C.: Hiding the presence of individuals from shared databases. In: SIGMOD, pp. 665–676 (2007)
21. NHS patient care data sharing scheme delayed. <http://www.theguardian.com/society/2014/dec/12/nhs-patient-care-data-sharing-scheme-delayed-2015-concerns> (2015). Accessed 6 Sept 2015
22. Picture archiving and communications system HSCIC. <http://systems.hscic.gov.uk/pacs> (2015). Accessed 6 Sept 2015
23. Population data bc. <https://www.popdata.bc.ca/data> (2015). Accessed 6 Sept 2015
24. Poulis, G., Loukides, G., Gkoulalas-Divanis, A., Skiadopoulos, S.: Anonymizing data with relational and transaction attributes. In: ECML/PKDD, pp. 353–369 (2013)
25. Samarati, P.: Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.* **13**(6), 1010–1027 (2001)
26. Sweeney, L.: K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **10**(5), 557–570 (2002)
27. Terrovitis, M., Liagouris, J., Mamoulis, N., Skiadopoulos, S.: Privacy preservation by disassociation. *Proc. VLDB* **5**(10), 944–955 (2012)
28. UK Data Protection Act. <http://www.legislation.gov.uk/ukpga/1998/29/contents> (2015). Accessed 6 Sept 2015
29. U.S. Veterans Health Administrations. <http://www.va.gov/health/> (2015). Accessed 6 Sept 2015
30. Wang, Q., Jin, H.: An analytical solution for consent management in patient privacy preservation. In: IHI, pp. 573–582 (2012)

Part I

Privacy in Data Sharing

Chapter 2

A Survey of Anonymization Algorithms for Electronic Health Records

Aris Gkoulalas-Divanis and Grigoris Loukides

Abstract Electronic Health Records (EHRs) contain various types of structured data about patients, such as patients' diagnoses, laboratory results, active medication, and allergies, which are increasingly shared to support a wide spectrum of medical analyses. To protect patient privacy, EHR data must be anonymized before their sharing. Anonymization ensures that the re-identification of patients and/or the inference of patients' sensitive information is prevented, and it is possible using several algorithms that have been proposed recently. In this chapter, we survey popular data anonymization algorithms for EHR data and explain their objectives, as well as the main aspects of their operation. After that, we present several promising directions for future research in this area.

2.1 Introduction

Electronic Health Records (EHRs) contain a wealth of clinical data, which include demographics, diagnosis codes, and medications. The data are typically used by physicians, for the diagnosis and treatment of diseases, and have been shown to improve patient care [23]. In addition, EHR data are becoming a valuable source for research [28]. For example, various analyses, such as controlling epidemics and predicting the risk of diabetes, can be applied to EHR data [60]. The large amounts of available EHR data and their usefulness in analytic tasks have led governments and major funders worldwide to support the wide sharing of EHR data [50, 53]. However, the sharing of EHR data must be performed in a way that offers patient anonymity [14]. Anonymization of EHR data is both a legal [62] and ethical requirement, as well as it is necessary in order to sufficiently protect the identity and sensitive information of patients.

A. Gkoulalas-Divanis (✉)
Smarter Cities Technology Centre, IBM Research, Dublin, Ireland
e-mail: arisdiva@ie.ibm.com

G. Loukides
School of Computer Science & Informatics, Cardiff University, Cardiff, UK
e-mail: g.loukides@cs.cf.ac.uk

Recently, the computer science and health informatics communities have proposed several algorithms for anonymizing EHR data. The main goal of these algorithms is to satisfy specific privacy requirements, while ensuring that the anonymized data remain useful for analysis. The privacy requirements are captured using formal *privacy models* and the enforcement of these models is performed by *data transformation*. In addition, data transformation must preserve *data utility*.

The majority of privacy models focus on blocking two important threats to patient privacy: (a) *identity disclosure*, which occurs when a patient is linked to their record in the published data [55, 57], and (b) *attribute disclosure* (or *sensitive information disclosure*) [49], which occurs when a patient is associated with sensitive information (e.g., HIV-positive status). Furthermore, data transformation is performed by: (a) *generalization*, which replaces values with more general, but semantically consistent values [44, 57], and (b) *suppression*, which deletes certain data values [55]. To preserve data utility, the anonymization algorithms aim to minimize the impact of data transformation. That is, the application of these algorithms should ensure that the anonymized EHR data: (a) retain important information, (b) remain useful in pre-specified medical data analysis tasks, and (c) satisfy certain requirements of data owners.

This chapter surveys anonymization algorithms for EHR data, with a focus on algorithms that are applied to patient demographics or to diagnosis codes. This is because most algorithms for EHR data aim to anonymize either of these two types of data, and they can be extended to deal with other patient attributes (e.g., the anonymization of diagnosis codes is similar to that of procedural codes [44]). To ease understanding and highlight the differences between anonymization algorithms, we first present a brief discussion of the main privacy models that are relevant to EHR data. Subsequently, we classify the existing privacy algorithms according to the privacy model they adopt. For each algorithm, we explain the main aspects of its operation, which include its privacy model, data transformation strategy, utility objective, and heuristic search strategy.

As it will become evident, there are major differences in the operation of the surveyed anonymization algorithms and also a number of important issues that remain outstanding. The chapter aims to shed light on the operation of these algorithms, to help data owners select the most appropriate algorithm for a specific EHR data publishing scenario. Furthermore, we aim to summarize outstanding issues related to the development of anonymization algorithms for EHR data, in order to motivate further research. For a more technical overview of a larger class of anonymization algorithms for medical data, we refer the readers to [20].

The remainder of this chapter is organized as follows. Section 2.2 presents privacy threats and models that have been proposed, and Sect. 2.3 reviews algorithms for anonymizing EHR data. Section 2.4 presents some interesting directions for future research, and Sect. 2.5 concludes this work.

2.2 Privacy Threats and Models

This section discusses the two major privacy threats that are relevant to the sharing of EHR data, namely *identity* and *attribute* disclosure. Next, it presents privacy models that aim to prevent each of these threats or both of them at the same time.

2.2.1 Privacy Threats

In the following, we classify the types of attributes that are found in EHR data, based on their role in the development of privacy models. Then, we discuss the privacy threats that arise in the sharing of EHR data.

Typically, there are three different types of attributes in structured EHR data: *direct identifiers*, *quasi-identifiers*, and *sensitive attributes*. Direct identifiers uniquely identify patients—examples of these attributes are names and phone numbers. Quasi-identifiers, on the other hand, are attributes which *in combination* identify a patient; quasi-identifying attributes include demographics [55, 65] and diagnosis codes [43]. Last, sensitive attributes are those that patients are not willing to be associated with. Examples of sensitive attributes are serious diseases, such as psychiatric diseases and forms of cancer.

Identity disclosure (or re-identification) [57, 65] occurs when a patient is associated with their record in the published EHR dataset. This is performed even when data are devoid of direct identifiers, since quasi-identifiers are contained in both the published dataset and in publicly available external data sources [55, 65]. For example, it has been estimated that over 87 % of U.S. citizens can be re-identified based on a combination of only three demographics (5-digit zip code, gender and date-of-birth) [55]. Many other identity disclosure incidents have also been reported [10]. For instance, in one attack students re-identified individuals in the Chicago homicide database by linking it with the social security death index.

Attribute disclosure (or sensitive information disclosure) [49] occurs when a patient is associated with information about their sensitive attributes. As discussed above, there are common types of attributes that are widely regarded as sensitive. The specification of sensitive attributes is generally left to data owners [34, 44, 49]. Although attribute disclosure has not transpired in the context of EHR data, it has raised serious privacy concerns, because the inference of sensitive medical information may cause financial and emotional embarrassment to patients.

2.2.2 Privacy Models

This section presents well-established privacy models that aim to protect data from attacks leading to identity and/or attribute disclosure. These include models that were specifically proposed for protecting EHR data, as well as others that can be easily applied for this purpose.

2.2.2.1 Models Against Identity Disclosure

In the following, we discuss well-established privacy models against identity disclosure, focusing on those that are applicable to demographics or diagnosis codes.

The most popular privacy model for protecting demographics is k -anonymity [55, 57]. k -anonymity requires each record in a dataset to contain the same values in the set of Quasi-IDentifier attributes (QIDs) with at least $k - 1$ other records in the dataset. k -anonymity protects from identity disclosure, because it limits the probability of attackers to successfully link an individual to his or her record, based on QIDs, to $1/k$. The parameter k controls the level of offered privacy and is set by data owners, usually to 3 or 5 in the context of EHR data [44].

A variation of k -anonymity is a privacy model, called k -map [56]. k -map considers that the linking is performed based on larger datasets (called *population tables*), from which the shared EHR dataset has been derived. Thus, k -map can be satisfied with lower data transformation than what is required by k -anonymity, but provides less protection. For example, identity disclosure may still occur when attackers know that a patient's record is included in the published dataset. Other models based on k -anonymity or k -map were proposed in [11] and in [16].

Several privacy models have been proposed to prevent identity disclosure attacks based on diagnosis codes. The work of He and Naughton [22] proposed *complete k -anonymity*, a model which assumes that *any* combination of diagnosis codes can lead to identity disclosure. This model is similar to k -anonymity. However, it may harm data utility unnecessarily because it is extremely difficult for attackers to know all the diagnoses in a patient record [43]. This led to the development of more flexible privacy models, such as k^m -*anonymity* [58]. k^m -anonymity uses a parameter m (typically $m = 2, \dots, 4$) to control the maximum number of diagnosis codes that may be known to an attacker, and it requires *each* combination of m diagnosis codes to appear in at least k records of the shared dataset.

Another flexible privacy model for diagnosis codes is *privacy-constrained anonymity* [44]. This model assumes that certain combinations of diagnosis codes may be known to an attacker and protects identity disclosure attacks, based on these combinations. By definition, privacy-constrained anonymity assumes that attackers know whether a patient's record is contained in the released dataset. This assumption is made by most research in the field (e.g., [22, 47, 58, 67]). Relaxing this assumption, however, is straightforward by following an approach similar to that of the k -map model, and can potentially offer more utility at the expense of privacy.

2.2.2.2 Models Against Attribute Disclosure

Privacy models against attribute disclosure can be classified according to the type of attributes they are applied to: (a) models for patient demographics, and (b) models for diagnosis codes. In what follows, we describe some representative privacy models from each of these groups.

The most popular privacy model that thwarts attribute disclosure attacks in patient demographics is l -diversity [49]. It requires each *anonymized group* in a dataset to contain at least l “well represented” (e.g., distinct or not “too” frequent) *sensitive attribute* values [49]. Other principles that thwart attribute disclosure, similarly to l -diversity, are (a,k) -*anonymity* [63] and p -*sensitive-k-anonymity* [61]. t -*closeness* [34] is another privacy model for protecting demographics from attribute disclosure. This model prevents an attacker from learning information about an individual’s sensitive value that is not available from the entire published EHR dataset. The models discussed so far prevent attribute disclosure based on categorical attributes (e.g., nationality). Privacy models to guard against the disclosure of sensitive ranges of values in numerical attributes have also been proposed [26, 32, 39, 41].

In addition, several privacy models have been proposed to protect attribute disclosure attacks when sharing diagnosis codes. For example, Cao et al. [5] proposed ρ -*uncertainty*, which limits the probability of associating an individual with any (single) diagnosis code to less than ρ . This model makes the (stringent) assumption that each diagnosis code in a patient record can be sensitive, and all the remaining codes in the record may be used for its inference. Another privacy model, called (h, k, p) -*coherence*, was proposed in [67]. The latter model prevents from both identity and sensitive information disclosure. It provides protection similar to that of both k^m -anonymity and l -diversity but for diagnosis codes.

A more recent privacy model for protecting privacy in the release of diagnosis codes is the *PS-rule based anonymity* model (*PS-rule* stands for *Privacy Sensitive rule*), which was proposed in [48]. This model thwarts both identity and sensitive information disclosure. The PS-rule based anonymity model offers significant benefits, as it allows data publishers to specify detailed privacy requirements, and it is more general than the models in [22, 58, 67].

2.3 Anonymization Algorithms

In this section, we provide a classification of algorithms that employ the privacy models in Sect. 2.2. These algorithms are summarized in Table 2.1. For each class of algorithms, we also discuss techniques that are employed in their operation.

2.3.1 Algorithms Against Identity Disclosure

As discussed in the introduction, anonymization algorithms transform data to enforce a privacy model in a way that preserves data utility. In the following, we briefly present approaches to transforming data and classify algorithms with respect to the strategies they adopt to transform data, their utility objectives, and their heuristic strategies. Based on this discussion, we subsequently present a detailed classification of the privacy algorithms.

Table 2.1 Anonymization algorithms that protect from identity and attribute disclosure (adapted from [20], with permission from Elsevier)

Privacy threat	Anonymization algorithms for	
	Demographics	Diagnosis codes
Identity disclosure	<i>k</i> -Minimal generalization [55]	
	OLA [9]	
	Incognito [29]	
	Genetic [25]	
	Mondrian [30, 31]	UGACLIP [44]
	TDS [12]	CBA [38]
	NNG [8]	UAR [37]
	Greedy [66]	Apriori [58]
	<i>k</i> -Member [4]	LRA [59]
	KACA [33]	VPA [59]
	Agglomerative [16]	mHgHs [36]
	(<i>k</i> , <i>k</i>)-anonymizer [16]	Recursive partition [22]
	Hilb [15]	
	iDist [15]	
	MDAV [7]	
	CBFS [27]	
Attribute disclosure	Incognito with <i>l</i> -diversity [49]	
	Incognito with <i>t</i> -closeness [34]	
	Incognito with (<i>a</i> , <i>k</i>)-anonymity [63]	Greedy [67]
	<i>p</i> -sensitive <i>k</i> -anonymity [61]	SuppressControl [5]
	Mondrian with <i>l</i> -diversity [64]	TDControl [5]
	Mondrian with <i>t</i> -closeness [35]	RBAT [46]
	Top Down [63]	Tree-based [48]
	Greedy algorithm [39]	Sample-based [48]
	Hilb with <i>l</i> -diversity [15]	
	iDist with <i>l</i> -diversity [15]	
	Anatomize [64]	

2.3.1.1 Data Transformation

The transformation of data is applied to quasi-identifiers, in order to prevent identity disclosure. To transform data, (a) microaggregation [6], (b) generalization [55], and (c) suppression [55] are applied. Microaggregation involves replacing a group of quasi-identifier values using a summary statistic (e.g., median) and is applicable to demographics but not to diagnosis codes. Generalization, on the other hand, suggests replacing quasi-identifier values by more general, but semantically consistent, values. Two generalization models, called *global* and *local* recoding, have been proposed [52]. Global recoding involves mapping the domain

of QIDs into generalized values, while local recoding involves mapping the values of individual records into generalized ones on a group-by-group basis. Finally, suppression involves the deletion of specific quasi-identifier values from the data, prior to their release.

2.3.1.2 Utility Objectives

As data transformation causes utility loss, anonymization algorithms aim at preserving data utility. This is achieved by applying one of the following general strategies: (a) quantifying information loss using measures, (b) assuming that data will be used in a specific data analysis task and aiming to preserve the accuracy of performing this task, and (c) satisfying pre-specified utility requirements (e.g., set by the data owners). These three strategies are discussed below.

The information loss measures that have been proposed to quantify data utility are used in cases when there is no specific use intended for the anonymized data. Accordingly, the goal of these measures is to introduce the minimum amount of overall distortion in the dataset to offer privacy. These measures are based on (a) the size of the formed anonymization groups, or (b) the characteristics of generalized values. Measures of the first category are based on the intuition that all records in an anonymization group are indistinguishable from one another, as they have the same value over QIDs. Thus, larger groups incur more information loss and smaller groups are preferred. Examples of measures that fall into the first category are the *Discernibility Metric* (DM) [3] and the *Normalized Average Equivalence Class Size* [30]. Examples of measures that belong to the second category are the *Generalization Cost* (GC) [1], the *Normalized Certainty Penalty* (NCP) [66], and the *Loss Metric* (LM) [25]. All of these measures are applicable to demographics. A recently proposed information-loss measure for diagnosis codes is *Information Loss Metric* (ILM) [38].

To quantify how well anonymized EHR data support a specific task, one can use the *Classification Metric* (CM), which was proposed in [25]. By optimizing this measure, anonymized data that helps building accurate classification models can be generated. The CM measure is expressed as the number of records whose class labels are different from that of the majority of records in their anonymized group, normalized by the dataset size. Furthermore, to support query answering using anonymized data, one can use the *Average Relative Error* (ARE) measure, which was proposed in [30]. ARE quantifies data utility by measuring the difference between the number of returned answers to a query using the anonymized and using the original data. Some measures for anonymizing data while supporting clustering applications were considered in [13]. Although these measures are also general, currently they have only been applied to patient demographics.

Several publishing scenarios involve the release of an anonymized dataset to support a specific medical study, or to data recipients having certain data analysis requirements. In such scenarios, knowledge of how the dataset will be analyzed can be exploited during anonymization to better preserve data utility. Samarati

proposed modeling data analysis requirements based on the minimum number of suppressed tuples, or on the height of hierarchies for categorical QID values [55]. Another approach was proposed by Xu et al. [66]. This approach prioritized the anonymization of certain attributes by using data-owner specified weights. To guarantee that the anonymized data will remain useful for the specified analysis requirements, Loukides et al. [42] proposed models for expressing data utility requirements. Utility requirements can be expressed at an attribute or at a value level and can be applied to patient demographics but not to diagnosis codes. Anonymizing diagnosis codes in a way that satisfies data utility requirements has been considered in [44]. The proposed approach models data utility requirements using sets of diagnosis codes, referred to as *utility constraints*. Collectively, utility constraints specify the information that the anonymized data should retain in order to be useful in intended medical analysis tasks, and form what is called a *utility policy*.

2.3.1.3 Heuristic Strategies

Anonymization algorithms employ heuristic search strategies to identify a “good” privacy solution. In what follows, we discuss different search strategies that have been applied to demographics and diagnosis codes.

Algorithms for demographics typically employ: (a) binary search on the lattice of possible generalizations [55], (b) a lattice search strategy similar in principle to the Apriori [2] algorithm, (c) genetic search on the lattice of possible generalizations [25], (d) data partitioning [24, 30], (e) data clustering [33, 39, 52, 66], or (f) space mapping [15]. The main idea behind strategies (a) to (c) is to represent the possible ways to generalize a value (in a quasi-identifier attribute) using a taxonomy, and then combine the taxonomies for all quasi-identifier attributes, to obtain a lattice. Thus, finding a way to generalize values can be performed by exploring the lattice using heuristics that avoid considering certain lattice nodes for efficiency reasons. The strategy (a) prunes the descendants of lattice nodes that are sufficient to satisfy a privacy model, while the strategies (b) and (c) prune lattice nodes that are likely to incur high utility loss. The latter nodes are identified while considering nodes that represent incrementally larger sets of generalized values, for strategy (a), or while selecting nodes by combining their descendants, as specified by a genetic algorithm, in the case of strategy (b).

More recent research has focused on developing methods that use strategies (d) and (e), which are applied to the records of a dataset and aim to organize records into carefully selected groups that help the preservation of privacy and the satisfaction of a utility objective. Both data partitioning and clustering-based strategies create groups iteratively, but they differ in the task they perform in an iteration. Specifically, partition-based strategies split records into groups, based on the value that these records have in a single quasi-identifier attribute, while clustering-based strategies merge two groups of records, based on the values of the records in all quasi-identifier attributes together.

Last, space mapping techniques [15] create a ranking of records, such that records with similar values in quasi-identifiers have similar ranks. Based on this ranking, groups of records are subsequently formed by considering a number of records that have consecutive ranks. Space mapping techniques achieve good effectiveness and efficiency.

Algorithms for diagnosis codes employ: (a) space partitioning in a bottom-up [44] or top-down [5] fashion, (b) space clustering [17], or (c) data partitioning in a top-down [22], vertical or horizontal [59] way.

Both strategies (a) and (b) attempt to find a set of generalized diagnosis codes that can be used to replace diagnosis codes in the original dataset. However, they differ in the way they operate. Specifically, space partitioning strategies require a taxonomy for diagnosis codes, while space clustering strategies lift this requirement and are more effective in terms of preserving data utility. On the other hand, data partitioning strategies are applied to transactions (records) instead of diagnosis codes, and they aim to create groups of transactions that can be subsequently anonymized with low data utility loss.

2.3.1.4 Classification of Algorithms

In what follows, a classification of algorithms for preventing identity disclosure is presented. The classification is based on the strategies that these algorithms adopt for (a) transforming quasi-identifiers, (b) preserving utility, and (c) heuristically searching the space for a “good” solution.

A. Algorithms for Demographics

Table 2.2 presents a classification of algorithms for preventing identity disclosure based on patients’ demographics. As can be seen, all algorithms adopt k -anonymity, with the exception of (k, k) -anonymizer [16] which adopts a similar model, called (k, k) -anonymity. Furthermore, most algorithms use generalization, except from (a) the algorithms in [29, 55], which use both generalization and suppression, and (b) the algorithms in [7, 27], which use microaggregation. Moreover, most algorithms aim at minimizing information loss and none of these algorithm takes into account specific utility requirements.

From the presented algorithms, the Genetic [25], Infogain Mondrian [31], and TDS [12] algorithms aim at releasing data in a way that allows for building accurate classifiers. The LSD Mondrian [31] algorithm is similar to Infogain Mondrian but uses a different utility objective measure, as its goal is to preserve the ability of using the released data for linear regression.

It can also be observed that several algorithms implement data partitioning heuristic strategies. Examples are [30, 31] that follow a top-down partitioning strategy, while the TDS algorithm [12] and the NNG [8] algorithm employ more sophisticated strategies to benefit data utility. On the other hand, the algorithms

Table 2.2 Algorithms that protect from identity disclosure based on demographics (Table adapted from [20], with permission from Elsevier)

Algorithm	Privacy model	Transformation	Utility objective	Heuristic strategy
k -Minimal generalization [55]	k -anonymity	gen and sup	Min. inf. loss	Binary lattice search
OLA [9]	k -anonymity	gen	Min. inf. loss	Binary lattice search
Incognito [29]	k -anonymity	gen and sup	Min. inf. loss	Apriori-like lattice search
Genetic [25]	k -anonymity	gen	Classification accuracy	Genetic search
Mondrian [30]	k -anonymity	gen	Min. inf. loss	Data partitioning
LSD Mondrian [31]	k -anonymity	gen	Regression accuracy	Data partitioning
Infogain Mondrian [31]	k -anonymity	gen	Classification accuracy	Data partitioning
TDS [12]	k -anonymity	gen	Classification accuracy	Data partitioning
NNG [8]	k -anonymity	gen	Min. inf. loss	Data partitioning
Greedy [66]	k -anonymity	gen	Min. inf. loss	Data clustering
k -Member [4]	k -anonymity	gen	Min. inf. loss	Data clustering
KACA [33]	k -anonymity	gen	Min. inf. loss	Data clustering
Agglomerative [16]	k -anonymity	gen	Min. inf. loss	Data clustering
(k,k)-anonymizer [16]	(k,k)-anonymity	gen	Min. inf. loss	Data clustering
Hilb [15]	k -anonymity	gen	Min. inf. loss	Space mapping
iDist [15]	k -anonymity	gen	Min. inf. loss	Space mapping
MDAV [7]	k -anonymity	mic	Min. inf. loss	Data clustering
CBFS [27]	k -anonymity	mic	Min. inf. loss	Data clustering

We use “gen”, “sup” and “mic” to refer to generalization, suppression, and microaggregation, respectively.

that employ clustering [4, 16, 51, 66] follow a similar greedy, bottom-up procedure, which aims at building clusters of at least k records by iteratively merging together smaller clusters of records, in a way that helps data utility preservation. The algorithms iHilb and iDist, both of which were proposed in [15], employ space partitioning strategies, while the algorithms in [9, 29, 55] use lattice-search strategies.

B. Algorithms for Diagnosis Codes

Algorithms for anonymizing diagnosis codes are summarized in Table 2.3. These algorithms adopt different privacy models, but they all use generalization and/or suppression. For example, the algorithms in [37, 38, 44] use suppression when generalization alone does not satisfy the specified utility constraints. In addition,

Table 2.3 Algorithms that protect against identity disclosure based on diagnosis codes (Table adapted from [20], with permission from Elsevier)

Algorithm	Privacy model	Transformation	Utility objective	Heuristic strategy
UGACLIP [44]	Privacy-constrained anonymity	gen and sup	Utility requirements	Bottom-up space partitioning
CBA [38]	Privacy-constrained anonymity	gen and sup	Utility requirements	Space clustering
UAR [37]	Privacy-constrained anonymity	gen and sup	Utility requirements	Space clustering
Apriori [58]	k^m -anonymity	gen	Min. inf. loss	Top-down space partitioning
LRA [59]	k^m -anonymity k^m -anonymity	gen	Min. inf. loss	Horizontal data partitioning
VPA [59]	k^m -anonymity k^m -anonymity	gen	Min. inf. loss	Vertical data partitioning
mHgHs [36]	k^m -anonymity	gen and sup	Min. inf. loss	Top-down space partitioning
Recursive partition [22]	Complete k -anonymity	gen	Min. inf. loss	Data partitioning

We use “gen” and “sup” to refer to generalization and suppression, respectively.

these algorithms aim at either satisfying utility requirements, or at minimizing information loss. The UGACLIP, CBA, and UAR algorithms adopt *utility constraints* to formulate utility requirements and attempt to satisfy them. However, these algorithms still favor solutions with low information loss, among those that satisfy the specified utility constraints. All other algorithms attempt to minimize information loss, which they quantify using two different measures.

Interestingly, all algorithms in Table 2.3 operate on either the space of diagnosis codes, or on that of the records of the dataset to be published. Specifically, UGACLIP [44] partitions the space of diagnosis codes in a bottom-up manner, whereas Apriori [58] and mHgHs [36] employ top-down partitioning strategies. Data partitioning strategies are employed by the Recursive partition [22], LRA [59] and VPA [59] algorithms.

2.3.1.5 Algorithms Against Attribute Disclosure

In the following, we present some background information for understanding the way algorithms for thwarting attribute disclosure work. Subsequently, we review some of the most popular algorithms that belong to this category.

The main goal of anonymization algorithms against attribute disclosure is to control the associations between quasi-identifier and sensitive values. Therefore, they create anonymous groups and merge them iteratively, until these associations are protected. This is possible using generalization and/or suppression. An alternative

technique called *bucketization* has also been proposed [64]. Bucketization works by releasing two projections of the dataset, one on quasi-identifiers and another on a sensitive attribute. Bucketization has been successfully applied to enforce l -diversity with low information loss in [64].

Many of the anonymization algorithms for protecting against attribute disclosure follow the same data transformation strategies and utility objectives, with the algorithms examined earlier. However, they additionally ensure that sensitive values are protected within each anonymized group. The approach of constructing anonymous data that prevent attribute disclosure, but are no more distorted than necessary, is referred to as *protection constrained*. Alternatively, data owners may want to produce anonymized data with a desired trade-off between data utility and privacy protection (against identity disclosure). This is possible using *trade-off constrained* algorithms [39–41]. These algorithms quantify and aim at optimizing the trade-off between the distortion and the protection against attribute disclosure.

A. Algorithms for Demographics

Table 2.4 presents a classification of algorithms for demographics. It can be easily seen that the majority of these algorithms follow the protection-constrained approach and employ generalization and/or suppression. The goal of most algorithms is to enforce l -diversity, t -closeness, p -sensitive k -anonymity, (a, k) -anonymity, or tuple-diversity. An exception is the Anatomize algorithm [64], which was specifically developed for enforcing l -diversity using bucketization. Note also that the algorithms in [39, 61, 63] are applied to both quasi-identifiers and sensitive attributes, and provide protection from both identity and attribute disclosure.

B. Algorithms for Diagnosis Codes

Algorithms for anonymizing diagnosis codes against attribute disclosure are summarized in Table 2.5. As can be seen, the algorithms adopt different privacy models, which are enforced using generalization and/or suppression. For example, TDControl [5] applies suppression when generalization alone cannot enforce ρ -uncertainty. In terms of heuristic search strategies, the algorithms in Table 2.5 employ a greedy search and operate on either the space of diagnosis codes, or on the transactions of the dataset to be published. For example, UGACLIP [44] partitions the space of diagnosis codes in a bottom-up manner, whereas Apriori [58] and mHgHs [36] employ top-down partitioning strategies.

Observe that all the algorithms in Table 2.5 operate on the space of diagnosis codes and either perform greedy search to discover diagnosis codes that can be suppressed with low data utility loss, or they employ space partitioning strategies. Specifically, TDControl, RBAT, and the Tree-based algorithm all employ top-down partitioning, while Sample-based uses both top-down and bottom-up partitioning strategies. The main difference between the strategy of TDControl and that of RBAT

Table 2.4 Algorithms for preventing attribute disclosure for demographics (Table adapted from [20], with permission from Elsevier)

Algorithm	Privacy model	Transformation	Approach	Heuristic strategy
Incognito with l -diversity [49]	l -diversity	Gen and sup	Protection constrained	Apriori-like lattice search
Incognito with t -closeness [34]	t -closeness	Gen and sup	Protection constrained	Apriori-like lattice search
Incognito with (a, k) -anonymity [63]	(a, k) -anonymity	Gen and sup	Protection constrained	Apriori-like lattice search
p -sens k -anon [61]	p -sensitive k -anonymity	Gen	Protection constrained	Apriori-like lattice search
Mondrian with [64]	l -diversity	Gen	Protection constrained	Data partitioning
l -diversity				
Mondrian with [35]	t -closeness	Gen	Protection constrained	Data partitioning
t -closeness				
Top Down [63]	(a, k) -anonymity	Gen	Protection constrained	Data partitioning
Greedy algorithm [39]	Tuple diversity		Trade-off constrained	Data clustering
		Gen and sup		
Hilb with l -diversity [15]	l -diversity	Gen	Protection constrained	Space mapping
iDist with l -diversity [15]	l -diversity	Gen	Protection constrained	Space mapping
Anatomize [64]	l -diversity	Buc	Protection constrained	Quasi-identifiers are released intact

We use “gen”, “sup”, and “buck” to refer to generalization, suppression, and bucketization, respectively.

is that the former is based on a taxonomy, which is used to organize diagnosis codes. This restricts the possible ways of partitioning diagnosis codes and may harm data utility unnecessarily. On the other hand, the Tree-based and Sample-based algorithms adopt a more flexible variation of the strategy employed in RBAT, which significantly enhances data utility.

2.4 Directions for Future Research

The sharing of EHR data is essential to support medical research in various areas, ranging from the detection of epidemics to the prediction of heart-related disorders. The sharing of EHR data, however, must be achieved while preserving patient privacy. The anonymization algorithms surveyed in this chapter make this possible. However, despite the progress in the development of anonymization algorithms for EHR data, there are several directions that warrant more investigation.

Table 2.5 Algorithms for preventing attribute disclosure based on diagnosis codes (Table adapted from [20], with permission from Elsevier)

Algorithm	Privacy model	Transformation	Approach	Heuristic strategy
Greedy [67]	(h, k, p) -coherence	Sup	Protection constrained	Greedy search
Suppress	ρ -uncertainty	Sup	Protection constrained	Greedy search
Control [5]				
TDControl [5]	ρ -uncertainty	Gen and sup	Protection constrained	Top-down space partitioning
RBAT [46]	PS-rule based anonymity	Gen	Protection constrained	Top-down space partitioning
Tree-based [48]	PS-rule based anonymity	Gen	Protection constrained	Top-down space partitioning
Sample-based [48]	PS-rule based anonymity	Gen	Protection constrained	Top-down and bottom-up space partitioning

We use “gen” and “sup” to refer to generalization and suppression, respectively.

First, EHR datasets become increasingly more complex and large, and this poses difficulties in anonymizing them using existing algorithms. For example, EHR datasets that contain both demographics and diagnosis codes, often need to be shared. The anonymization of these datasets, while preserving data utility, is challenging because: (a) the two attribute types (demographics and diagnosis codes) cannot be anonymized separately, and (b) it is not straightforward to preserve data utility when anonymizing both of the attribute types together.

This creates the need for sophisticated data anonymization algorithms that are able to deal with the problem, while maintaining the utility of EHR data in practice. A first step towards solving this problem has been made recently [54]. However, the algorithms proposed in [54] are not designed for EHR data. As another example, consider EHR datasets that are several GBs in size, such as those used to support longitudinal studies. These datasets do not fit into the main memory of a typical computer, and thus they cannot be anonymized using the algorithms that are surveyed in this chapter. It would be worthwhile to develop scalable anonymization algorithms for handling such large EHR datasets.

Second, while the focus of this chapter has been on data anonymization, there are other privacy threats that are relevant to EHR data sharing. An example is the risks stemming from the mining of published data [18, 19, 21]. These risks are posed because mining reveals knowledge that apply to a large number of patients, which is not hidden by anonymization algorithms. To mitigate these risks, sensitive knowledge patterns need to be detected and concealed from the data, so that they cannot be discovered by data mining techniques applied on the shared EHR data.

Last, we note that the chapter focused on technical means (and more specifically algorithms) for protecting EHR data and considered only data held by a single party (data owner). However, non-technical means, including key legal frameworks (e.g., [62]) and various data sharing regulations are also important to protect data.

Furthermore, there are many cases in which EHR data are contributed by multiple parties (e.g., various healthcare institutions that collaboratively perform a study). The anonymization of data in this setting [45] is quite challenging, since it requires: (a) combining different parts of data in a secure manner, and (b) taking into account changes in data, as well as the possibility of collusion between attackers, which makes privacy preservation more challenging.

2.5 Conclusion

This chapter presented a survey of data anonymization algorithms for protecting patient demographics and diagnosis codes in the context of Electronic Health Record (EHR) data. We classified and discussed many popular data anonymization algorithms, which can be used to prevent the main privacy threats that arise in EHR data publishing. We also explained the main aspects of these algorithms, including the privacy models they follow, the data transformation strategies they apply, the utility objectives they aim to satisfy, and heuristic search strategies they use to discover “good” solutions. In addition, we provided a discussion of several research directions that in our opinion warrant further investigation.

Acknowledgements Grigoris Loukides is partly supported by a Research Fellowship from the Royal Academy of Engineering.

References

1. Aggarwal, G., Kenthapadi, F., Motwani, K., Panigrahy, R., Thomas, D., Zhu, A.: Approximation algorithms for k-anonymity. *J. Privacy Technol.* **3**, 1–8 (2005)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: VLDB, pp. 487–499 (1994)
3. Bayardo, R., Agrawal, R.: Data privacy through optimal k-anonymization. In: 21st ICDE, pp. 217–228 (2005)
4. Byun, J., Kamra, A., Bertino, E., Li, N.: Efficient k-anonymization using clustering techniques. In: DASFAA, pp. 188–200 (2007)
5. Cao, J., Karras, P., Raïssi, C., Tan, K.: *rho*-uncertainty: inference-proof transaction anonymization. *Proc. VLDB* **3**(1), 1033–1044 (2010)
6. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. Knowl. Data Eng.* **14**(1), 189–201 (2002)
7. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Min. Knowl. Disc.* **11**(2), 195–212 (2005)
8. Du, Y., Xia, T., Tao, Y., Zhang, D., Zhu, F.: On multidimensional k-anonymity with local recoding generalization. In: ICDE ’07, pp. 1422–1424 (2007)
9. El Emam, K., Dankar, F.K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J.P., Walker, M., Chowdhury, S., Vaillancourt, R., Roffey, T., Bottomley, J.: A globally optimal k-anonymity method for the de-identification of health data. *J. Am. Med. Inform. Assoc.* **16**(5), 670–682 (2009). doi: [10.1197/jamia.M3144](https://doi.org/10.1197/jamia.M3144)

10. El Emam, K., Jonker, E., Arbuckle, L., Malin, B.: A systematic review of re-identification attacks on health data. *PLoS ONE* **6**(12), e28,071 (2011). <http://dx.doi.org/10.1371%2Fjournal.pone.0028071>
11. Emam, K.E., Dankar, F.K.: Protecting privacy using k-anonymity. *J. Am. Med. Inform. Assoc.* **15**(5), 627–637 (2008)
12. Fung, B.C.M., Wang, K., Yu, P.S.: Top-down specialization for information and privacy preservation. In: *ICDE*, pp. 205–216 (2005)
13. Fung, B.C.M., Wang, K., Wang, L., Hung, P.C.K.: Privacy-preserving data publishing for cluster analysis. *Data Knowl. Eng.* **68**(6), 552–575 (2009)
14. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: a survey on recent developments. *ACM Comput. Surv.* **42**, 1–53 (2010)
15. Ghinita, G., Karras, P., Kalnis, P., Mamoulis, N.: Fast data anonymization with low information loss. In: *Proceedings of the 33rd International Conference on Very large Data bases, VLDB '07*, pp. 758–769 (2007)
16. Gionis, A., Mazza, A., Tassa, T.: k-anonymization revisited. In: *ICDE*, pp. 744–753 (2008)
17. Gkoulalas-Divanis, A., Loukides, G.: PCTA: privacy-constrained clustering-based transaction data anonymization. In: *EDBT PAIS*, p. 5 (2011)
18. Gkoulalas-Divanis, A., Loukides, G.: Revisiting sequential pattern hiding to enhance utility. In: *KDD*, pp. 1316–1324 (2011)
19. Gkoulalas-Divanis, A., Verykios, V.S.: Hiding sensitive knowledge without side effects. *Knowl. Inf. Syst.* **20**(3), 263–299 (2009)
20. Gkoulalas-Divanis, A., Loukides, G., Sun, J.: Publishing data from electronic health records while preserving privacy: a survey of algorithms. *J. Biomed. Inform.* **50**(1), 4–19 (2014)
21. Gwadera, R., Gkoulalas-Divanis, A., Loukides, G.: Permutation-based sequential pattern hiding. In: *IEEE International Conference on Data Mining (ICDM)*, pp. 241–250 (2013)
22. He, Y., Naughton, J.F.: Anonymization of set-valued data via top-down, local generalization. *Proc. VLDB* **2**(1), 934–945 (2009)
23. Hsiao, C., Hing, E.: Use and characteristics of electronic health record systems among office-based physician practices: United states, 2001–2012. In: *NCHS Data Brief*, pp. 1–8 (2012)
24. Iwuchukwu, T., Naughton, J.F.: K-anonymization as spatial indexing: Toward scalable and incremental anonymization. In: *VLDB*, pp. 746–757 (2007)
25. Iyengar, V.S.: Transforming data to satisfy privacy constraints. In: *KDD*, pp. 279–288 (2002)
26. Koudas, N., Zhang, Q., Srivastava, D., Yu, T.: Aggregate query answering on anonymized tables. In: *ICDE '07*, pp. 116–125 (2007)
27. Laszlo, M., Mukherjee, S.: Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Trans. Knowl. Data Eng.* **17**(7), 902–911 (2005)
28. Lau, E., Mowat, F., Kelsh, M., Legg, J., Engel-Nitz, N., Watson, H., Collins, H., Nordyke, R., Whyte, J.: Use of electronic medical records (EMR) for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data. *Clin. Epidemiol.* **3**(1), 259–272 (2011)
29. LeFevre, K., DeWitt, D., Ramakrishnan, R.: Incognito: efficient full-domain k-anonymity. In: *SIGMOD*, pp. 49–60 (2005)
30. LeFevre, K., DeWitt, D., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: *ICDE*, p. 25 (2006)
31. LeFevre, K., DeWitt, D., Ramakrishnan, R.: Workload-aware anonymization. In: *KDD*, pp. 277–286 (2006)
32. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Workload-aware anonymization techniques for large-scale datasets. *ACM Trans. Database Syst.* **33**(3), 1–47 (2008)
33. Li, J., Wong, R., Fu, A., Pei, J.: Achieving -anonymity by clustering in attribute hierarchical structures. In: *DaWaK*, pp. 405–416 (2006)
34. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l-diversity. In: *ICDE*, pp. 106–115 (2007)
35. Li, N., Li, T., Venkatasubramanian, S.: Closeness: A new privacy measure for data publishing. *IEEE Trans. Knowl. Data Eng.* **22**(7), 943–956 (2010)

36. Liu, J., Wang, K.: Anonymizing transaction data by integrating suppression and generalization. In: Proceedings of the 14th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining, PAKDD '10, pp. 171–180 (2010)
37. Loukides, G., Gkoulalas-Divanis, A.: Utility-preserving transaction data anonymization with low information loss. *Expert Syst. Appl.* **39**(10), 9764–9777 (2012)
38. Loukides, G., Gkoulalas-Divanis, A.: Utility-aware anonymization of diagnosis codes. *IEEE J. Biomed. Health Inform.* **17**(1), 60–70 (2013)
39. Loukides, G., Shao, J.: Capturing data usefulness and privacy protection in k-anonymisation. In: SAC, pp. 370–374 (2007)
40. Loukides, G., Shao, J.: An efficient clustering algorithm for -anonymisation. *J. Comput. Sci. Technol.* **23**(2), 188–202 (2008)
41. Loukides, G., Shao, J.: Preventing range disclosure in k-anonymised data. *Expert Syst. Appl.* **38**(4), 4559–4574 (2011)
42. Loukides, G., Tziatzios, A., Shao, J.: Towards preference-constrained -anonymisation. In: DASFAA International Workshop on Privacy-Preserving Data Analysis (PPDA), pp. 231–245 (2009)
43. Loukides, G., Denny, J., Malin, B.: The disclosure of diagnosis codes can breach research participants' privacy. *J. Am. Med. Inform. Assoc.* **17**, 322–327 (2010)
44. Loukides, G., Gkoulalas-Divanis, A., Malin, B.: Anonymization of electronic medical records for validating genome-wide association studies. *Proc. Natl. Acad. Sci.* **17**(107), 7898–7903 (2010)
45. Loukides, G., Gkoulalas-Divanis, A., Malin, B.: An integrative framework for anonymizing clinical and genomic data, Chap. 8. In: Database Technology for Life Sciences and Medicine, pp. 65–89. World Scientific, Singapore (2010)
46. Loukides, G., Gkoulalas-Divanis, A., Shao, J.: Anonymizing transaction data to eliminate sensitive inferences. In: DEXA, pp. 400–415 (2010)
47. Loukides, G., Gkoulalas-Divanis, A., Malin, B.: COAT: Constraint-based anonymization of transactions. *Knowl. Inf. Syst.* **28**(2), 251–282 (2011)
48. Loukides, G., Gkoulalas-Divanis, A., Shao, J.: Efficient and flexible anonymization of transaction data. *Knowl. Inf. Syst.* **36**(1), 153–210 (2013)
49. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: ICDE, p. 24 (2006)
50. Mailman, M., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., et al.: The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181–1186 (2007)
51. Massimo, R., Angiulli, F., Pizzuti, C.: Descry: a density based clustering algorithm for very large dataset. In: 5th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'04), pp. 25–27 (2004)
52. Nergiz, M.E., Clifton, C.: Thoughts on k-anonymization. *Data Knowl. Eng.* **63**(3), 622–645 (2007)
53. Ollier, W., Sprosen, T., Peakman, T.: UK biobank: from concept to reality. *Pharmacogenomics* **6**(6), 639–646 (2005)
54. Poulis, G., Loukides, G., Gkoulalas-Divanis, A., Skiadopoulos, S.: Anonymizing data with relational and transaction attributes. In: ECML/PKDD (3), pp. 353–369 (2013)
55. Samarati, P.: Protecting respondents identities in microdata release. *IEEE Trans. Knowl. Data Eng.* **13**(9), 1010–1027 (2001)
56. Sweeney, L.A.: Computational disclosure control: a primer on data privacy protection. Ph.D. thesis (2001). AAI0803469
57. Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Uncertainty Fuzziness Knowledge Based Syst.* **10**, 557–570 (2002)
58. Terrovitis, M., Mamoulis, N., Kalnis, P.: Privacy-preserving anonymization of set-valued data. *Proc. VLDB* **1**(1), 115–125 (2008)
59. Terrovitis, M., Mamoulis, N., Kalnis, P.: Local and global recoding methods for anonymizing set-valued data. *VLDB J.* **20**(1), 83–106 (2011)

60. Tildesley, M.J., House, T.A., Bruhn, M., Curry, R., O'Neil, M., Allpress, J., Smith, G., Keeling, M.: Impact of spatial clustering on disease transmission and optimal control. *Proc. Natl. Acad. Sci.* **107**(3), 1041–1046 (2010)
61. Truta, T., Vinay, B.: Privacy protection: p-sensitive k-anonymity property. In: ICDE Workshops, p. 94 (2006)
62. U.S. Department of Health and Human Services Office for Civil Rights: HIPAA administrative statute and rules, <http://www.hhs.gov/ocr/privacy/hipaa/administrative/> (September 6, 2015)
63. Wong, R.C., Li, J., Fu, A., K.Wang: alpha-k-anonymity: An enhanced k-anonymity model for privacy-preserving data publishing. In: KDD, pp. 754–759 (2006)
64. Xiao, X., Tao, Y.: Anatomy: simple and effective privacy preservation. In: VLDB, pp. 139–150 (2006)
65. Xiao, X., Tao, Y.: Personalized privacy preservation. In: SIGMOD, pp. 229–240 (2006)
66. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.W.C.: Utility-based anonymization using local recoding. In: KDD, pp. 785–790 (2006)
67. Xu, Y., Wang, K., Fu, A.W.C., Yu, P.S.: Anonymizing transaction databases for publication. In: KDD, pp. 767–775 (2008)

Chapter 3

Differentially Private Histogram and Synthetic Data Publication

Haoran Li, Li Xiong, and Xiaoqian Jiang

Abstract Differential privacy has recently emerged as one of the strongest privacy guarantees by making few assumptions on the background or external knowledge of an attacker. Differentially private data analysis and publishing have received considerable attention in biomedical communities as promising approaches for sharing medical and health data, while preserving the privacy of individuals represented in data records. In this chapter, we provide a broad survey of the recent works in differentially private histogram and synthetic data publishing. We categorize most recent and emerging techniques in this field from two major aspects: (a) various data types (e.g., relational data, transaction data, dynamic stream data, etc.), and (b) parametric, and non-parametric techniques. We also present some challenges and future research directions for releasing differentially private histogram and synthetic data in health and medical data.

3.1 Introduction

The problem of preserving patient privacy in disseminated biomedical datasets has attracted increasing attention by both the biomedical informatics and computer science communities. The goal is to share a “sanitized” version of the individual records (microdata) that simultaneously provides utility for data users and privacy protection for the individuals represented in the records. In the biomedical domain, many text de-identification tools focus on extracting identifiers from different types of medical documents, and use simple identifier removal or replacements according to the HIPAA safe harbor method [32] for de-identification. Several studies and reviews have evaluated the re-identification risk of linking de-identified data by the HIPAA safe harbor method with external data, such as voter registration lists. Many

H. Li (✉) • L. Xiong

Department of Mathematics & Computer Science, Emory University, Atlanta, GA, USA
e-mail: hli57@emory.edu; lxiong@emory.edu

X. Jiang

Division of Biomedical Informatics, University of California, San Diego, La Jolla, CA, USA
e-mail: x1jiang@ucsd.edu

studies have proposed or applied formal anonymization methods on medical data. While anonymization is still the dominant approach in practice, the main limitation of microdata release with de-identification is that it often relies on assumptions of certain background or external knowledge (e.g., availability of voter registration lists) and only protects against specific attacks (e.g., linking or re-identification attacks) [25].

Differential privacy has emerged as a promising privacy mechanism under which a “sanitized” version of individual records (microdata) can be generated and published. This privacy principle makes few assumptions on the background or external knowledge of an attacker, and thus provides a strong provable privacy guarantee. A statistical aggregation or computation satisfies ϵ -differential privacy, if the outcomes are formally “indistinguishable” (“indistinguishable” is formally and quantitatively defined in Dwork [12]): the outcome probability differs by no more than a multiplicative factor e^ϵ) when run with and without any particular record in the dataset, where ϵ is a privacy parameter that limits the maximum amount of influence a record can have on the outcome.

In this chapter, we first introduce the concept of differential privacy, as well as mechanisms for achieving differential privacy. Then, we summarize a number of state-of-the-art data release techniques under differential privacy in the database community, while categorizing existing methods based on the type of data they handle: relational data, transaction data, and dynamic stream data. There are many methods designed for learning specific models or patterns with differential privacy, but for the purposes of this work we only focus on histogram and synthetic data generation, which are useful for exploratory analysis. Finally, we investigate potential challenges and future directions for applying differentially private data release techniques to healthcare and biomedical data.

3.2 Differential Privacy

In this section, we first introduce the fundamental concept of differential privacy, and then discuss mechanisms of achieving differential privacy. Finally, we present composition theorems for a sequence of private mechanisms to guarantee overall differential privacy.

3.2.1 Concept of Differential Privacy

Differential privacy has emerged as one of the strongest privacy definitions for statistical data release. It guarantees that if an adversary knows complete information of all the tuples in D except one, the output of a differentially private randomized algorithm should not give the adversary too much additional information about the remaining tuples. We say that datasets D and D' are two neighboring datasets when

D and D' differ in only one tuple t , and write that $\|D - D'\| = 1$. Let us make it clear that there exist two scenarios: the *replacement scenario* and the *insertion/deletion scenario*. In the replacement scenario, $\|D - D'\| = 1$ means that we can obtain D' by changing only one tuple t from D and tuple t has different values in the two datasets (i.e. $|D| = |D'|$ ¹); in the insertion/deletion scenario, it means that we can obtain D' by removing or adding only one tuple from D and tuple t is present in only one of the two datasets (i.e. $|D| - |D'| = 1$). For concreteness, in this paper we consistently use the second definition. A formal definition of differential privacy is given as follows:

Definition 3.1 (ϵ -Differential Privacy [9]). Let \mathcal{A} be a randomized algorithm over two neighboring datasets D and D' , and let \mathcal{O} be any arbitrary set of possible outputs of \mathcal{A} . Algorithm \mathcal{A} satisfies ϵ -differential privacy if and only if the following holds:

$$\Pr[\mathcal{A}(D) \in \mathcal{O}] \leq e^\epsilon \Pr[\mathcal{A}(D') \in \mathcal{O}]$$

Intuitively, differential privacy ensures that the released output distribution of \mathcal{A} remains nearly the same whether or not an individual tuple is in the dataset.

3.2.2 Mechanisms of Achieving Differential Privacy

Laplace Mechanism The most common mechanism to achieve differential privacy is the Laplace mechanism [9] that adds a small amount of independent noise to the output of a numeric function f to fulfill ϵ -differential privacy of releasing f , where the noise is drawn from *Laplace distribution* with a probability density function $\Pr[\eta = x] = \frac{1}{2b} e^{-\frac{|x|}{b}}$. A Laplace noise has a variance $2b^2$ with a magnitude of b . The magnitude b of the noise depends on the concept of *sensitivity* that is defined as follows.

Definition 3.2 (Sensitivity [9]). Let f denote a numeric function and the sensitivity of f is defined as the maximal L_1 -norm distance between the outputs of f over the two data sets D and D' which differs in only one tuple. Formally,

$$\Delta_f = \max_{D,D'} \|f(D) - f(D')\|_1.$$

With the concept of sensitivity, the noise follows a zero-mean Laplace distribution with the magnitude $b = \frac{\Delta_f}{\epsilon}$. To fulfill ϵ -differential privacy for a numeric function f over D , it is sufficient to publish $f(D) + X$, where X is drawn from $\text{Lap}(\frac{\Delta_f}{\epsilon})$.

Geometric Mechanism The geometric mechanism [18] adds integer noise, and is usually preferred for queries q with integer answers, e.g., COUNT queries [7]. When $f(D)$ is integer-valued, X can be drawn from the geometric distribution

¹ $|D|$ means the data cardinality or the number of tuples in database D .

$$\Pr[X = x] = \frac{1 + \alpha}{1 - \alpha} \alpha^{|x|}$$

where $\alpha = e^{-\epsilon/\Delta_f}$, $x \in \mathcal{Z}$.

Exponential Mechanism McSherry and Talwar [28] propose the exponential mechanism by selecting an output $t \in \mathcal{T}$ that is close to the optimum with respect to a utility function under differential privacy.

Theorem 3.1 (Exponential Mechanism [28]). *For an input data set D , output range \mathcal{T} , privacy parameter ϵ , and a utility function $q : D \times \mathcal{T} \rightarrow \mathcal{R}$, a randomized mechanism that chooses an output $t \in \mathcal{T}$ with probability proportional to $\exp(\frac{\epsilon q(D, t)}{2\Delta_q})$ guarantees ϵ -differential privacy.*

The exponential mechanism takes as inputs a dataset D , output range \mathcal{T} , privacy parameter ϵ , and a utility function $q : D \times \mathcal{T} \rightarrow \mathcal{R}$ that assigns a real valued score to every output $t \in \mathcal{T}$, where higher scores have better utility. Let $\Delta_q = \max_{r, D, D'} |q(D, t) - q(D', t)|$ be the sensitivity of the utility function, which means the maximum difference in q over two input data sets that differ in only a single individual, for all r . The probability associated with each output t increases proportionally to $\exp(\frac{\epsilon q(D, t)}{2\Delta_q})$; that is, the output with a higher score is exponentially more likely to be sampled. So the mechanism samples an output t by building a probability distribution over the output range \mathcal{T} with an exponential mechanism of the utility function. Here, the exponential mechanism significantly bias the distribution and tends to favor outputs with high scores [28].

The motivation behind the exponential mechanism is that it is useful in situations in which we only want to select the top-K “best” responses from a domain, but adding noise directly to the computed quantity can completely destroy its value (such as χ^2 statistics are computed from a set of SNPs) and directly adding Laplace noise could dramatically distort the original values because of high sensitivity.

Staircase Mechanism The staircase mechanism, proposed by Geng et al. [17], is an optimal ϵ -differentially private mechanism for single real-valued query functions. The objective here is to minimize the worst case cost among all possible query outputs with differential privacy as the constraint. The probability density function of the noise in the staircase mechanism is staircase-shaped, symmetric around the origin, monotonically decreasing and geometrically decaying. The staircase mechanism can be viewed as a geometric mixture of uniform probability distributions, providing a simple algorithmic description for the mechanism. It has been shown in [17] that the staircase mechanism performs better than the Laplace mechanism for single real-valued query functions.

3.2.3 Composition Theorems

For a sequence of differentially private mechanisms, the composition [27] theorems guarantee the overall privacy.

Theorem 3.2 (Sequential Composition [27]). *For a sequence of n mechanisms M_1, \dots, M_n and each M_i provides ϵ_i -differential privacy, the sequence of M_i provides $(\sum_{i=1}^n \epsilon_i)$ -differential privacy.*

Theorem 3.3 (Parallel Composition [27]). *If D_i are disjoint subsets of the original database and M_i provides α -differential privacy for each D_i , then the sequence of M_i provides α -differential privacy.*

3.3 Relational Data

In this section, we first present the problem setting of histogram and synthetic data publication for relational data under differential privacy. Then, we classify various methods into three categories: *parametric algorithms*, *semi-parametric algorithms* and *non-parametric algorithms*, and describe all methods in each subsection.

3.3.1 Problem Setting

Running Example Assume that we have a relational dataset with attributes `age` and `income`, and a two-dimensional count data cube or histogram.² The domain values of `age` are $20 \sim 30$, $30 \sim 40$ and $40 \sim 50$; the domain values of `income` are $0 \sim 10K$, $10K \sim 20K$ and $> 20K$. Each cell in the histogram bin represents the population count corresponding to the `age` and `income` values.

Differentially private histogram publication means releasing a differentially private histogram with noisy histogram bin counts, as shown in Fig. 3.1. The left noisy histogram of Fig. 3.1 employs a direct Laplace mechanism. The right noisy histogram first applies advanced sub-histogram partitioning techniques, and then injects independent Laplace noise to each sub-histogram. We will demonstrate this noisy histogram generation process in the following subsections. Note that in the problem of differentially private histogram publication, histogram bin counts may be negative since a random Laplace noise is sampled from a Laplace distribution with mean zero.

The main idea behind differentially private synthetic data generation is to build a private statistical model from the data and then to sample points from the model.

²Data cube and histogram are the same objects in this chapter.

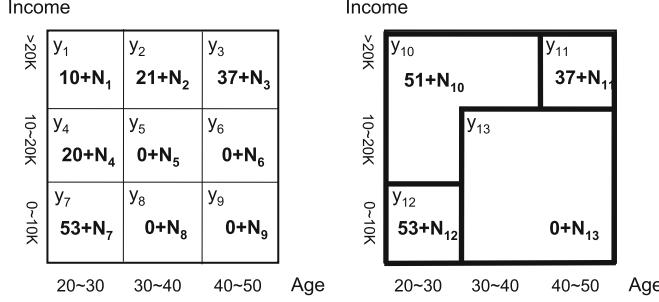


Fig. 3.1 Example: released cell histogram (*left*) and subcube histogram (*right*), and N_i is a random Laplace noise (see Sect. 3.2 for Laplace mechanism)

These sampled points form the synthetic data, which is then released instead of the original data. The private synthetic data can be generated from private histograms after rounding all negative bin counts to zero.

Utility Measures Range count queries are often used to measure the utility of the private histograms and synthetic data. Random range count queries with random query predicates covering all attributes, are defined in the following:

Select COUNT(*) from D

Where $A_1 \in I_1$ and $A_2 \in I_2$ and \dots and $A_m \in I_m$.

For each attribute A_i , I_i is a random interval generated from the domain of A_i . The query accuracy is primarily measured by the relative error defined as follows: for a query q , $A_{act}(q)$ is the true answer to q on the original data. $A_{noisy}(q)$ denotes the noisy answer to q when using differentially private histogram and synthetic data publication methods. Then the relative error (RE) is defined as:

$$RE(q) = \frac{A_{noisy}(q) - A_{act}(q)}{\max\{A_{act}(q), s\}}$$

where s is a sanity bound to mitigate the effects of queries with extremely small query answers. Given a workload of queries, the utility over n random count queries can be calculated as

$$ARE(q) = \frac{1}{n} \sum_{i=1}^n \frac{A_{noisy}(q) - A_{act}(q)}{\max\{A_{act}(q), s\}}$$

Other measures, such as KL-divergence and classification accuracy are also used (interested readers can see [1, 29]).

Three Categories of Existing Methods The main approaches of existing work can be classified into three categories:

1. Parametric methods that fit the original data to a multivariate distribution and make inferences about the parameters of the distribution while preserving differential privacy, as shown in Fig. 3.2. Synthetic data is then generated from the private multivariate distribution.
2. Non-parametric methods that learn empirical distributions from the data through histograms, as shown in Fig. 3.3;
3. Semi-parametric methods that capture the dependence implicit in the high-dimensional datasets, while using non-parametric methods to generate a marginal histogram of each dimension, as shown in Fig. 3.4.

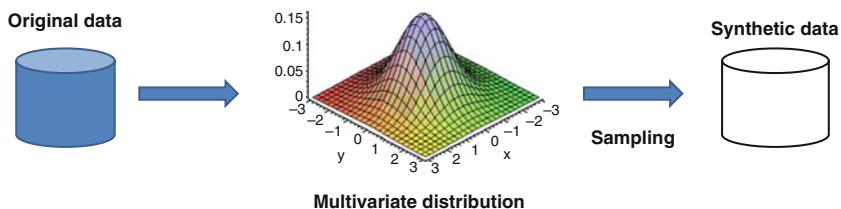


Fig. 3.2 Generate synthetic data via parametric methods

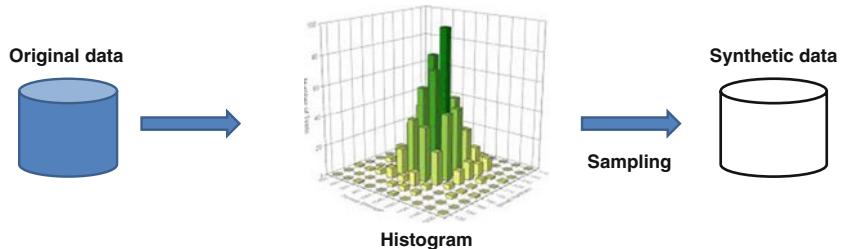


Fig. 3.3 Generate synthetic data via non-parametric methods

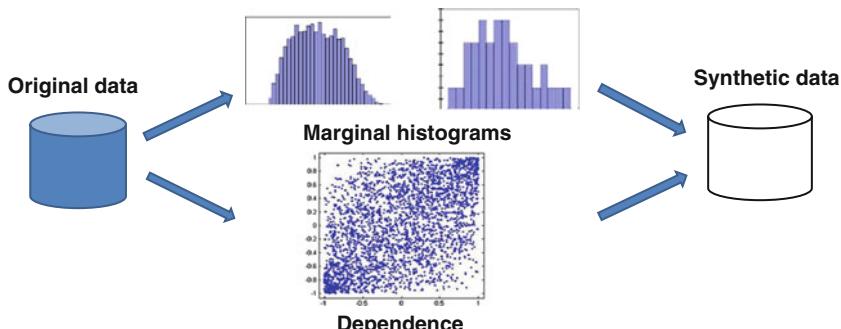


Fig. 3.4 Generate synthetic data via semi-parametric methods

3.3.2 Parametric Algorithms

Machanavajjhala et al. [26] propose a differentially private Multinomial-Dirichlet synthesizer (MD) to release the commuting patterns of the population of the United States. They build a multinomial distribution with Dirichlet prior estimation as an initial mechanism to fit the original data. As the original data is extremely sparse and the resulting synthetic data contains a large number of extremely unlikely events, the paper proposes an (ϵ, δ) -probabilistic differential privacy mechanism, a relaxed and probabilistic version of differential privacy. With the new definition, MD revises the initial mechanism to throw away useless synthetic data, and minimizes the negative effects of the sparse data and large domain in the map application. Since MD uses a multivariate distribution to fit the original data and samples the synthetic data from this distribution, we categorize MD to parametric algorithms.

3.3.3 Semi-parametric Algorithms

DPCopula Li et al. [24] propose a DPCopula framework to generate high dimensional and large domain DP synthetic data. DPCopula computes a differentially private copula function³ and samples direct synthetic data from the function that effectively captures the dependence implicit in the high-dimensional datasets. With the copula functions, DPCopula can separately consider the margins and the joint dependence structure of the original data, instead of modeling the joint distribution or empirical histogram of all dimensions, as shown in Fig. 3.4.

Step 1: Create a differentially private marginal histogram for each dimension to obtain DP empirical marginal distribution.

Step 2: Use DP MLE (Maximum Likelihood Estimation) or DP kendall to estimate the DP correlation matrix \tilde{P} , and each entry of \tilde{P} is a correlation coefficient of each pairwise attributes.

Step 3: Sample DP synthetic dataset from the private copula function.

We call methods like DPCopula *semi-parametric*, because they compute private marginal histograms for all dimensions, which belongs to non-parametric techniques, and model the joint dependence by estimating the private correlation matrix, which can be considered as parametric technique.

³In linguistics, a copula is a word used to link the subject of a sentence with a predicate, such as the word “is” in the sentence “The sky is blue”. In probability and statistics theory, a copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform.

3.3.4 Non-parametric Algorithms

LPA The method by Dwork et al. [14] publishes a DP histogram by adding independent Laplace random noise to each cell count of the original histogram, as shown in Fig. 3.1 (left). We call this method LPA, which lays the foundation for other, more advanced, methods and is considered as a baseline strategy. While LPA works well for low-dimensional data, it becomes problematic for high dimensional and large domain data, because the original count in each histogram bin will be extremely small, leading to large perturbation error incurred by Laplace random noise.

Fourier Transform Barak et al. [2] uses LPA to obtain a DP frequency matrix, then transforms it to the Fourier domain and adds Laplace noise in this domain. With the noisy Fourier coefficients, it employs linear programming to create a non-negative frequency matrix. The problem is that (a) post-processing is not shown to improve accuracy, and (b) the method requires solving a linear program, where the number of variables equals to the number of entries in the frequency matrix. This can be computationally challenging for practical datasets with large domain space. For instance, for the eight-dimensional datasets, each with domain size of 1000, the number of variables is: $\prod_{i=1}^m |A_i| = 10^{24}$.

Privelet+ Xiao et al. [35] propose a Privelet method by applying a wavelet transform on the original histogram, then adding polylogarithmic noise to the transformed data. Privelet takes the histogram H of a relational table T and a differentially private version H' of H . The general steps are as follows:

First, it applies a wavelet transform on H . Generally speaking, a wavelet transform is an invertible linear function, i.e., it maps H to another matrix C , such that (a) each entry in C is a linear combination of the entries in H , and (b) H can be losslessly reconstructed from C . The entries in C are referred to as the wavelet coefficients. Wavelet transforms are traditionally defined for ordinal data, and a special extension for nominal data is investigated (see [35] for details).

Second, Privelet adds independent Laplace noise to each wavelet coefficient in a way that ensures ϵ -differential privacy. This results in a new matrix C' with noisy coefficients.

In the third step, Privelet (optionally) refines C' and subsequently maps C' back to a noisy frequency matrix H' , which is returned as the output. The refinement of C' may arbitrarily modify C' , but it does not utilize any information from the original table T or H . In other words, this third step of Privelet depends only on C' . This ensures that Privelet does not leak any information about T , except for what has been disclosed in C' .

A Privelet+ method is also proposed to handle tables with small discrete attributes. It first divides the original histogram into sub-histograms along the dimensions specified by small discrete attributes, then applies Privelet on each sub-histogram, and finally assembles all noisy sub-histograms into one histogram.

NoiseFirst and StructureFirst Xu et al. [36] propose NoiseFirst and StructureFirst techniques to release differentially private histograms. The main difference of these two techniques lies in the relative order of the noise injection and the histogram structure computation steps. NoiseFirst first uses LPA to generate a noisy histogram \tilde{H} , then runs the dynamic programming histogram construction algorithm on \tilde{H} . The NoiseFirst algorithm can be interpreted as a post-optimization technique to the baseline LPA method, by merging adjacent noisy counts. The intuitive idea behind NoiseFirst is that averaging over the neighboring noisy counts is able to eliminate the impact of zero-mean Laplace noise, based on the large number theorem. Since NoiseFirst fails to exploit the reduced sensitivity after bin merging, StructureFirst is developed which calculates the histogram structure on the original data before adding the noise on the histogram bin structure. It first constructs the optimal histogram by choosing boundaries between neighbouring bins with exponential mechanism, then adds Laplace noise on the average counts of all sub-histogram partitions. NoiseFirst and StructureFirst are mainly applied on perturbing histograms with low-dimensional original data for its high computation complexity.

Filtering and Sampling Cormode et al. [7] developed a series of filtering and sampling techniques to obtain a compact histogram summary H'' of sparse data under differential privacy, given a histogram H of original relation table T . The size of H'' (i.e., the number of histogram bins) can be much smaller than H . The first and simplest summarization method is the *high-pass filter*.

For every non-zero histogram bin count $H(i)$, it adds independent geometric random noise with parameter α to get a noisy count $\tilde{H}(i)$ which is added to H'' if $\tilde{H}(i)$ is larger than a threshold θ . For zero histogram bin counts, uniformly at random selects k locations i from H such that $H(i) = 0$, where k is sampled from a binomial distribution with parameter $\frac{2\alpha^\theta}{1+\alpha}$. For each of these k locations, it draws a value with a coin flipping sign from $Pr[X \leq x] = (1 - \alpha)^{x-\theta+1}$ and adds it to H'' . It has been proved that the high-pass filter generates a summary with the same distribution as the baseline LPA method for $\theta \geq 1$.

The other two advanced methods are based on *threshold sampling* and *priority sampling*. They add each non-zero histogram bin count to H'' due to a probability or priority weight, k is sampled from a binomial distribution with an alternative parameter form, and the value of zero bin count is sampled from an alternative distribution. The limitation is that if a large number of small-count non-zero entries exists in the histogram, it will give zero entries a higher probability to be in the final summary, leading to less accurate summary results. In addition, it needs careful selection of the appropriate values for several parameters, including sample size and filter threshold ([7] did not provide a principled approach to determine them).

DPCube DPCube uses a two-phase partitioning strategy. First, *a cell based partitioning based on the domains* (not the data) is used to generate a fine-grained equi-width cell histogram (as in the baseline LPA strategy), which gives an approximation of the original data distribution. Then a synthetic dataset D_c is released based on the cell histogram. Second, *a multi-dimensional partitioning*

based on kd-tree is performed on D_c to obtain uniform or close to uniform partitions. The resulted partitioning keys are used to partition the original database and obtain a noisy count for each of the partitions, resulting in a v-optimal histogram. Finally, given a user-issued query, an estimation component uses either the v-optimal histogram or both histograms to compute an answer of the query.

DPCube preserves differential privacy because the second multi-dimensional partitioning step accesses only the private synthetic dataset D_c with no impact on the privacy guarantee. The second step uses a kd-tree based space partitioning strategy that seeks to produce close to uniform partitions in order to minimize the estimation error within a partition. It starts from the root node which covers the entire space. At each step, a splitting dimension and a split value from the range of the current partition on that dimension are chosen heuristically to divide the space into subspaces. The algorithm repeats until a pre-defined requirement are met. Here, several metrics, such as information entropy and variance, can be used to measure the uniformity of a partition and make the decision whether to split the current partition and to select the best splitting point.

PSD Cormode et al. [6] design a class of differentially private spatial decompositions (PSDs). These partition the space into smaller regions, and report statistics on the points within each region. Queries can then be answered by intersecting the query region with the decomposition. Query answers are expected to be more accurate by performing a spatial decomposition which results in compact regions with sufficiently many points and a more uniform distribution. Spatial decompositions involved by PSDs include *data-independent* tree structures, such as quadtrees which recursively divide the data space into equal quadrants, and *data-dependent* trees, such as the popular kd-trees which aim to better capture the data distribution by using original spatial data.

The simplest PSD can be formed by instantiating a data-independent tree such as a full quadtree, but computing counts for each node via the Laplace mechanism. See Fig. 3.5 for an example. For data-independent trees, releasing the structure of the index (i.e., the node rectangles) does not endanger the privacy of any individual

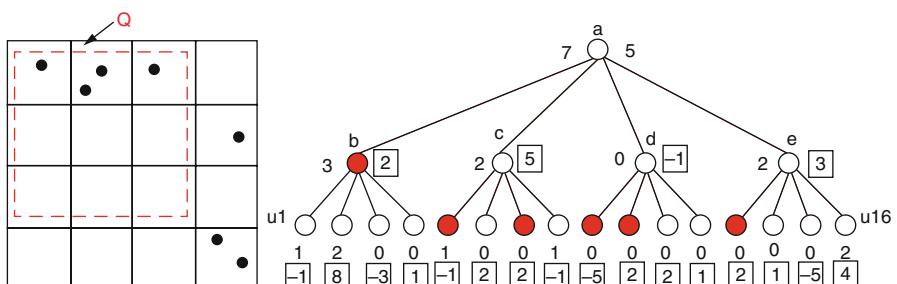


Fig. 3.5 Example of private quadtree: noisy counts (inside boxes) are released; actual counts, although depicted, are not released. Query Q (dotted red rectangle) could be answered by adding noisy counts of marked nodes (Color figure online) [6]

since the concept of sensitivity (Definition 3.2) captures the maximum change in the output caused by the presence or absence of a single tuple. For data-dependent trees, a private tree-structure needs to be built since the choice of splits is based on the true data, and potentially reveals information. This problem can be induced to compute a differentially private median for a given set of spatial points. Four methods for computing a private median are referred as follows.

1. **Smooth sensitivity** tailors the noise to be more specific to the set of multi-dimensional points whose values are in non-decreasing order for some certain dimension. However, smooth sensitivity has a slightly weaker privacy guarantee that is (ϵ, δ) -differential privacy (see [31] for more details).
2. **Exponential mechanism** [28] (Theorem 3.2.1) is a general method as an alternative to the Laplace mechanism: instead of adding random noise to the true value, the output value is drawn from a probability distribution over all possible outputs, so that the differential privacy condition is satisfied. Applied to median, the exponential mechanism EM returns x with $\Pr[EM(C) = x] \propto \exp(-\frac{\epsilon}{2}|\text{rank}(x) - \text{rank}(x_m)|)$, where C is a (multi)set of n values in non-decreasing order in some domain range, x_m is its median value and $\text{rank}(x)$ denote the rank of x in C .
3. **Cell-based Method** is a heuristic proposed in DPCube. It imposes a fixed resolution grid over C and then computes the median based on the noisy counts in the grid cells. When applied to a hierarchical decomposition, a fixed grid is computed over the entire data, then medians are computed from the subset of grid cells in each node. Cell counts have sensitivity 1. The accuracy of this method depends on the coarseness of the grid relative to the data distribution.
4. **Noisy Mean** is a heuristic from [19], which replaces median with mean. A private mean can be computed by computing a noisy sum (with sensitivity M) and a noisy count (with sensitivity 1), and outputting their ratio. If the count is reasonably large, this is a fair approximation to the mean, though there is no guarantee that this is close to the median. One heuristic of noisy mean is that the original data should have approximately uniform distribution, because the mean and median differ significantly for skewed data distributions (e.g., the age of patients with Alzheimer's disease).

EFPA Acs et al. [1] study an Enhanced Discrete Fourier Transform (EFPA) mechanism for generating differentially private histograms. It applies the Fourier transform to a histogram and compresses it by removing high-frequency components using the exponential mechanism. They improve the performance of FPA by designing a more accurate score function for the exponential mechanism and exploiting the intrinsic correlation among the Fourier coefficients of real-valued histograms. The general steps of EFPA are as follows:

Step 1: Compute the DFT coefficients $F = DFT^{real}(H)$ of a given histogram H with length n by discrete Fourier transform. The length of F is m , and $m = \frac{n+1}{2}$.

Step 2: Remove the last $m - k$ coefficients from F , which correspond to the high-frequency components in H , whereas the first k elements of F , denoted by F^k ,

preserve the low frequencies in H , and therefore represent the high-level trends of H . Note that k is selected by exponential mechanism, with the probability proportional to $\exp(-\frac{\epsilon U(H,k)}{\sqrt{2}})$, where $U(H,k) = \sqrt{\sum_{i=k+1}^n |F_{i-1}|^2 + \frac{2z}{\epsilon}}$, and $z = 2k + 1$.

Step 3: Generate the noisy version of F^k , denoted by \tilde{F}^k , by LPA: add i.i.d Laplace noise $\text{Lap}(\frac{2\sqrt{z}}{\epsilon})$ to each coefficient in F^k , where $z = 2k + 1$.

Step 4: Pad \tilde{F}^k to be a m -dimensional vector by appending $m - k$ zeros, which is denoted by $PAD^m(\tilde{F}^k)$.

Step 5: The inverse DFT is applied to $PAD^m(\tilde{F}^k)$ to obtain a noisy version of H .

The total differential privacy budget ϵ is uniformly divided between LPA (Step 3) and exponential mechanism (Step 2), therefore, EFPA is ϵ -differentially private due to the sequential composition theorem. EFPA is an advanced version of Discrete Fourier Transform (DFT).

P-HP The authors in [1] also propose P-HPartition that uses a divisible hierarchical clustering (partitioning) scheme to compress histograms. The intuition is that histogram bins belonging to the same cluster have similar counts, and hence can be approximated by their mean value (i.e., cluster center). It is often sufficient to release only the noisy cluster centers, which have a smaller sensitivity.

The challenge is how to make hierarchical partitioning of histogram bins under differential privacy. P-HP leverages on a tree structure of the clustering hierarchy. The initial partition containing all n bins forms the root of the tree; each bisection splits a partition (represented as a node in the tree) into two child nodes. Each bisection consists of two steps:

1. Select a partition who has not been bisected for more than d times before, which is made data-independent and therefore does not violate differential privacy.
2. Bisect the selected partition using exponential mechanism.

In order to make sure that the selection of a partition to bisect is data-independent, P-HP maintains a queue of all partitions, and always picks up the first partition in the queue to bisect. After each bisection, the resultant sub-partitions are appended to the end of the queue, and the new configuration (i.e., all partitions in the queue) is saved.

DiffGen The DiffGen method [29] releases a differentially private synthetic data table especially for classification analysis. The input relational table $T(A_1^{pr}, \dots, A_d^{pr}, A^{cls})$ includes two categories of attributes:

1. A class attribute A^{cls} that contains the class value, and the goal of the data miner is to build a classifier to accurately predict the value of this attribute.
2. A set of d predictor attributes $(A_1^{pr}, \dots, A_d^{pr})$, whose values are used to predict the class attribute.

The class attribute is categorical, and the predictor attribute can be either numerical or categorical, while for each categorical-predictor attribute A_i^{pr} , a taxonomy tree is provided (see Fig. 3.6 for a taxonomy tree example).

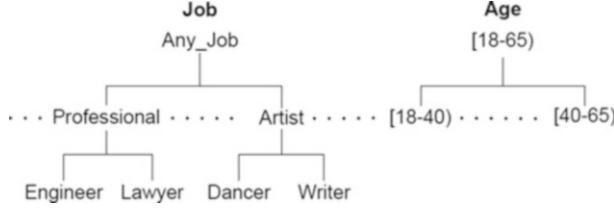


Fig. 3.6 Taxonomy tree of attributes [29]

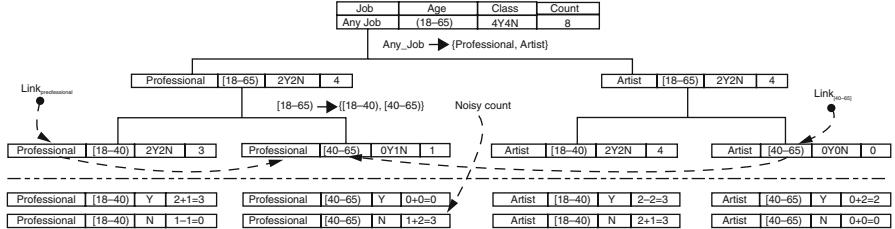


Fig. 3.7 Tree for partitioning records [29]

DiffGen first generalizes the predictor attributes $(A_1^{pr}, \dots, A_d^{pr})$ and divides the original table into several equivalence groups, where all the records within a group have the same attribute values. Then, the algorithm publishes the noisy counts of the groups. The general idea is to anonymize the original data by a sequence of partitions, starting from the topmost general state as shown in Fig. 3.7. Then it pushes down the partition iteratively by specializing an attribute and some value in the current partition. At each iteration, DiffGen probabilistically selects a candidate attribute for partition. Candidates are selected using the exponential mechanism with their score values, and different heuristics (e.g., information gain) can be used to determine the score function of the candidates. Then the score of the affected candidates is updated due to the partition. Finally, Laplace noise is injected to the original count of each partition group. DiffGen terminates after a given number of partitions.

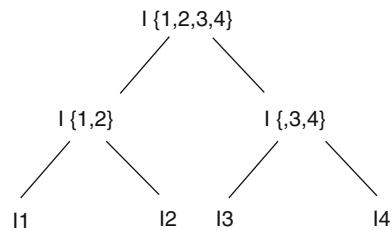
3.4 Transaction Data

In this section, we first present the problem setting for the case of transaction/set-valued data under differential privacy, and then explain a set of methods that have been proposed to solve it.

Table 3.1 A sample set-valued dataset [5]

TID	Items
t_1	$\{I_1, I_2, I_3, I_4\}$
t_2	$\{I_2, I_4\}$
t_3	$\{I_2\}$
t_4	$\{I_1, I_2\}$
t_5	$\{I_2\}$
t_6	$\{I_1\}$
t_7	$\{I_1, I_2, I_3, I_4\}$
t_8	$\{I_2, I_3, I_4\}$

Fig. 3.8 A context-free taxonomy tree of the sample data in Table 3.1 [5]



3.4.1 Problem Setting

Let $I = \{I_1, I_2, \dots, I_{|I|}\}$ be the universe of items, where $|I|$ is the size of the universe. The multiset $D = \{t_1, t_2, \dots, t_{|D|}\}$ denotes a set-valued or transaction dataset, where each record $t_i \in D$ is a non-empty subset of I . Table 3.1 presents an example of a set-valued dataset with the item universe $I = \{I_1, I_2, I_3, I_4\}$.

3.4.2 DiffPart

Chen et al. [5] proposed DiffPart, a differentially private sanitization algorithm that recursively partitions a given set-valued dataset, based on a context-free taxonomy tree. A context-free taxonomy tree is a taxonomy tree, whose internal nodes are a set of their leaves, not necessarily the semantic generalization of the leaves. For example, Fig. 3.8 presents a context-free taxonomy tree for Table 3.1, and one of its internal nodes $I_{\{1,2,3,4\}} = \{I_1, I_2, I_3, I_4\}$. We say that an item can be generalized to a taxonomy tree node if it is in the nodes set. For example, I_1 can be generalized to $I_{\{1,2\}}$ because $I_1 \in \{I_1, I_2\}$.

DiffPart takes as inputs the raw set-valued dataset D , the fan-out f used to construct the taxonomy tree, and also the total privacy budget ϵ , specified by the data publisher, and returns a sanitized dataset \tilde{D} satisfying ϵ -differential privacy. It starts by creating the context-free taxonomy tree, and then generalizes all records to a single partition with a common representation. The common representation is called the *hierarchy cut*, consisting of a set of taxonomy tree nodes. It recursively

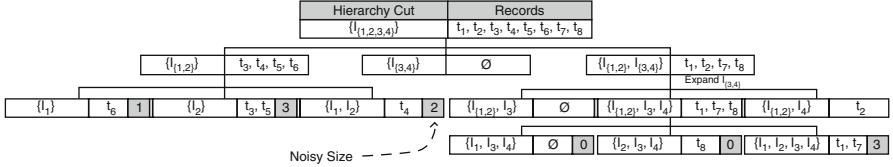


Fig. 3.9 The partitioning process of Fig. 3.1 [5]

distributes the records into disjoint sub-partitions with more specific representations in a top-down manner based on the taxonomy tree. For each sub-partition, DiffPart determines if it is empty in a noisy way and further splits the sub-partitions considered “non-empty”. The method stops when no further partitioning is possible in any sub-partition. Given the dataset in Table 3.1 and a fan-out value 2, a possible partitioning process is illustrated in Fig. 3.9.

3.4.3 Private FIM Algorithms

Bhaskar et al. [3] proposes two mechanisms for choosing the top- K most frequent patterns in a sensitive transaction dataset. The first one consists of applying K rounds of the exponential mechanisms. In each round, a set of itemsets with fixed size are sampled. For a given dataset, the score of an item set is a truncated version of its frequency, and the analysis of privacy is determined by bounding the sensitivity of the truncated frequency. The second mechanism is to add independent Laplace noise to the frequencies of all itemsets and choose the top- K item sets with the highest perturbed frequencies. The authors show that it suffices to add noise only $O(\frac{K}{\epsilon})$.

3.4.4 PrivBasis

Li et al. [23] proposes a PrivBasis algorithm by avoiding the selection of top k itemsets from a very large candidate set. The method first computes the number of size-one items λ , which are involved in the top- K frequent item sets. Second, it obtains the λ most frequent items F among all item sets. It can be approximately achieved that F includes exactly the items that appear in the top- K itemsets. Subsequently, it computes a set of the most frequent pairs of items P among F . The desired goal is that P includes exactly the pairs of items that appear in the top- K item sets. Finally, it constructs a θ -basis set B with small width and length, and obtains noisy frequencies of item sets in a given candidate set whose size is reduced based on B . Top- K item sets can be selected.

3.5 Stream Data

In this section, we present the problem setting for the case of stream data under differential privacy, and discuss methods that have been proposed to solve it.

3.5.1 Problem Setting

An aggregate time series $\mathbf{X} = \{x_k\}$ is defined as a set of values of a variable x observed at a discrete time point k , with $0 \leq k < T$, where T is the length of the series. In practical applications, \mathbf{X} may often be an aggregate count series, such as the daily number of patients diagnosed of Influenza, or the hourly count of drivers passing by a gas station. If each element in \mathbf{X} is the value of a multivariate variable or a histogram, i.e., $\mathbf{X} = \{H_k\}$, then we define $\mathbf{X} = \{H_k\}$ as a dynamic dataset or histogram data. In this section, both aggregate time series and dynamic datasets are considered as dynamic stream data.

User-Level Privacy vs. Event-Level Privacy There exists two definitions of differential privacy in the private release of stream data, namely *event-level* privacy and *user-level* privacy. The former protects any single event at any time point, while the latter preserves all the events of any user over the entire time points. In user-level privacy, the neighbouring datasets are defined as two series of dynamic datasets differing in only one user.

The two works that initiated the study on this setting are [11, 13], which consider the input data stream as a bitstring, and at each time point release the number of 1's seen so far. Both works focus on event-level privacy and release the number of event occurrences (i.e., a counter) at every time point. In this scenario, the effect of a single event at some time point spans over all subsequent publications. The work in [11] adds the noisy update value with Laplace noise of magnitude $\frac{1}{\epsilon}$ to the counter as an update arrives at the next time point. This method performs well when the 1's arrive frequently over the entire stream, but may be ineffective under sparse appearance of 1's. To improve this limitation, [11] also studies that the publication of a new counter may be postponed until a predefined number of 1's has arrived.

Dwork et al. [13] proposes a binary tree method to improve the efficiency of [11]. Based on a stream with fixed length N , they dynamically construct a full binary tree as the updates of the stream arrive. Each node stores a noisy value with scale logarithmic in N . Then, at each update i , they first identify the subtree it belongs to, and report the current counter after adding the noisy values stored in the roots of these subtrees.

A similar idea is proposed in [4]. The authors build a full binary tree on the updates, where each node stores the noisy sum of the updates in its subtree with scale logarithmic in N . At the i th update, the algorithm identifies the maximal subtrees covering updates 1 to i , and publishes the sum of the values contained in their roots. The authors also extend their method to handle infinite streams, by building a new binary tree after 2^k updates arrive, where k is a positive integer.

In this section, we focus on recent methods under user-level privacy. We first introduce methods under finite streams, then investigate w-event privacy, which combines user-level and event-level privacy to bridge the gap between infinite and finite stream data. The mechanism is simple for event-level privacy, since it is sufficient to sample private statistics at every time stamp with ϵ -privacy budget, and the overall private time-series statistics still guarantees ϵ -differential privacy.

A Direct Extension of LPA A baseline solution to sharing differentially private time series is to apply the standard Laplace perturbation at each time stamp. If every query satisfies $\frac{\epsilon}{N}$ -differential privacy, based on the sequential composition theorem the sequence of queries guarantees ϵ -differential privacy. The problem is that the utility of released stream data will be reduced as N increases, and the solution can not be applied to infinite streams.

3.5.2 Discrete Fourier Transform

Rastogi and Nath [33] propose a discrete Fourier transform (DFT) method. EFPA is an advanced version of DFT. Note that DFT is “a one-time” release of time-series data, meaning that we need to collect original time-series data at all time stamps, and then apply DFT to the overall time series. We describe DFT as follows:

Step 1: Compute the DFT coefficients $F = DFT(H)$ of a times-series X with length T by discrete Fourier transform. The length of F is also T . The j -th coefficient is given as: $DFT(X)_j = \sum_{i=0}^{T-1} \exp\left(\frac{2\pi\sqrt{-1}}{T}ji\right)x_i$.

Step 2: Remove the last $m - k$ coefficients from F , which correspond to the high-frequency components in X , whereas the first k elements of F , denoted by F^k , preserve the low frequencies in X , and therefore represent the high-level trends of X . Note that k is an input to the algorithm.

Step 3: Generate the noisy version of F^k , denoted by \tilde{F}^k , using the Laplace perturbation mechanism: add i.i.d Laplace noise $Lap\left(\frac{\sqrt{k}}{\epsilon}\right)$ to each coefficient in F^k .

Step 4: Pad \tilde{F}^k to be a n -dimensional vector by appending $n - k$ zeros, which is denoted by $PAD^n(\tilde{F}^k)$.

Step 5: The inverse DFT (IDFT) is applied to $PAD^n(\tilde{F}^k)$ to obtain a noisy version of X . The j -th element of the inverse is: $\frac{1}{T} \sum_{i=0}^{T-1} \exp\left(-\frac{2\pi\sqrt{-1}}{T}ji\right)x_i$.

3.5.3 FAST

Fan et al. [16] proposed a FAST framework to release time-series data with differential privacy (illustrated in Fig. 3.10). We present the outline of FAST as follows:

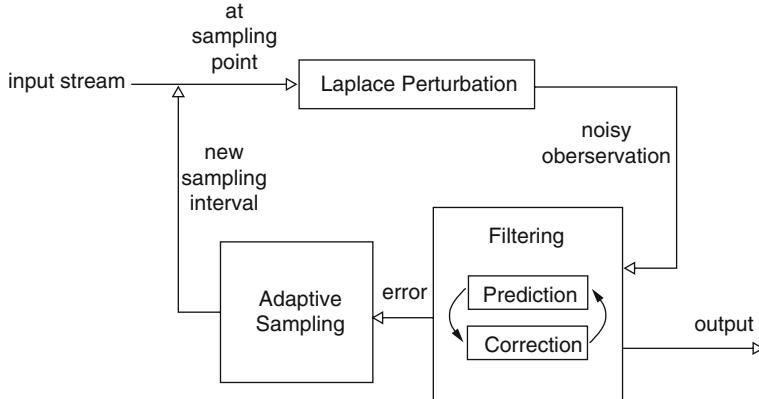


Fig. 3.10 The FAST framework [16]

- For each time stamp k , the adaptive sampling component determines whether to sample/query the input time-series or not.
- If k is a sampling point, the data value at k is perturbed by the Laplace mechanism to guarantee ϵ_1 -differential privacy, where $\epsilon_1 = \frac{\epsilon}{M}$, and M is the maximum number of samples FAST allows for a time series, ϵ is the privacy budget.
- The filtering component produces a prediction of a data value based on an internal state model at every time stamp. The prediction, i.e., prior estimate, is released to output at a non-sampling point, while a posterior estimate, i.e., correction of the noisy observation and prediction, is released at a sampling point.
- The error between the prior and the posterior is then fed through the adaptive sampling component to adjust the sampling rate. Once the user-specified privacy budget ϵ is used up, the system will stop sampling the input series.

There are three important characteristics of FAST. First, it requires that the original time-series data follows a linear process model, meaning that the states at consecutive time stamps can be modeled by $x_k = x_{k-1} + \omega$, where ω is a white Gaussian noise. Second, it allows for fully automated adaptation to changing data dynamics. Finally, FAST supports real-time time-series data publishing.

3.5.4 *w*-Event Privacy

w-event privacy [21] is proposed as an extension of differential privacy to address release of infinite streams, by combining user-level and event-level privacy. It protects any event sequence occurring within any window of w timestamps. It is event-privacy with $w = 1$ and converges to user-level privacy with $w = \infty$. *w*-event privacy guarantees user-level ϵ -differential privacy for every sub sequence of length w (or over w timestamps) *anywhere* (i.e., it can start from any timestamp) in the

original series of dynamic datasets. w -neighboring series of dynamic datasets, \mathbf{D}_w , and $\hat{\mathbf{D}}_w$, can be defined as the user-level neighbors under any sub-sequence of length w , anywhere. w -event privacy can be formally defined as follows:

Definition 3.3 (w -Event ϵ -Differential Privacy). Let \mathcal{A} be a randomized mechanism over two w -neighboring series of dynamic datasets \mathbf{D}_w , and $\hat{\mathbf{D}}_w$, and let \mathcal{O} be any arbitrary set of possible outputs of \mathcal{A} . Algorithm \mathcal{A} satisfies w -event ϵ -differential privacy (or, w -event privacy) if and only if it holds that:

$$\Pr[\mathcal{A}(\mathbf{D}_w) \in \mathcal{O}] \leq e^\epsilon \Pr[\mathcal{A}(\tilde{\mathbf{D}}_w) \in \mathcal{O}]$$

The authors in [21] also proposed two mechanisms, Budget Distribution (BD) and Budget Absorption (BA), to allocate the budget within one w -timestamp window. They propose a sampling approach which computes the noisy distance between the dataset at the current time point and the original dataset at the latest sampling point, and then compares the noisy distance with the perturbation noise to be added if current dataset is to be released. If the distance is greater than the perturbation noise, a noisy dataset is released at current time stamp. The perturbation noise is determined by their privacy budget allocation schemes, Budget Distribution (BD) and Budget Absorption (BA), that allocate the budget to different timestamps in the w -event window. Specifically, BD allocates the privacy budget in an exponentially decreasing fashion, in which earlier timestamps obtain exponentially more budget than later ones. BA starts by uniformly distributing the budget to all w timestamps, and accumulates the budget of non-sampling timestamps, which can be allocated later to the sampling timestamps. A main drawback of their approach is that the privacy budget may be exhausted prematurely (sampling too frequently in the beginning) or not fully utilized during all w timestamps (sampling not frequent enough), leading to suboptimal utility of the released data.

3.6 Challenges and Future Directions

In this section, we present a series of challenges related to differentially private data releases in the biomedical domain and discuss some future directions. There are a number of characteristics that need to be addressed in exploring and applying differentially private techniques for healthcare data. The introduction of new privacy preserving methods would have to consider these issues before being employed widely. The aspects below are summarized from previous works and our experiences in the healthcare data field.

3.6.1 Variety of Data Types

Health data contains a large number of categorical data and numerical data, especially continuous numeric data with high precision (e.g., blood pressure, weight, height). This may disable existing techniques to be employed directly in health data. For example, most state-of-the-art differentially private data release methods are based on count queries (e.g., computing general or partial histograms) and the addition of Laplace noise to numeric continuous data can distort the original data largely. Hierarchical methods are often used to generalize continuous numeric data by existing DP techniques. However, numerical attributes of health data sometimes can not be generalized, for example, the dose of a certain drug which should be strictly controlled.

3.6.2 High Dimensionality

Healthcare data is often high dimensional, contains many data errors, and does not follow a certain distribution. The utility of released synthetic data will be deteriorated as the number of dimensions increase. How to generate private synthetic data for high dimensional data effectively and efficiently is still an open topic in the community of differential privacy. The correlation and general joint distribution of raw data may not be retained sufficiently by the private data, especially in the case of high-dimensional data with unexpected distributions. Since most of the existing methods are based on count queries, good utility of count queries does not necessarily mean accurate simulation of correlation and joint distribution. The granularities of count queries are difficult to determine.

3.6.3 Correlated Constraints Among Attributes

Many attributes in healthcare datasets are correlated or have natural constraints. For example, one drug may precede another drug for treatment, or a number of drugs are taken or never taken in combination. Correlations may also exist between drugs and diagnoses, and lab results and diagnoses.

Independent privacy-preserving perturbation may reduce the trust of data analysts on the perturbed data, which is a barrier to the wide acceptability and adoption of differentially private techniques. For example, if the perturbed data shows two drugs that are known to interact in a way that can be damaging to a patient's health, or a drug that would never be prescribed with a particular treatment appearance for the same patient, or a dose that does not make sense for a patient, then data analysts will cease to trust the released private data. In practice, this requires higher accuracy of synthetic data generation and creates challenges for existing methods that add noise to original data.

3.6.4 Limitations of Differential Privacy

Setting and Explaining the Privacy Budget ϵ Setting an appropriate value for ϵ is a challenging task and it is still an open topic even in differential privacy literature. It is difficult for common data holders to measure how much protection the private synthetic data will provide for a specific ϵ value. For example, if a user wants to know the number of smokers who had lung cancer in an original dataset, then what does $\epsilon = 0.1$ mean? Is this enough to protect the appearance of one individual smoker with lung cancer in the raw dataset? Does the value of ϵ depend on the dataset itself or the domain universe of attributes?

To date, there are no theoretical analysis or experimental evaluation on how to appropriately choose a value of ϵ . Dwork considers the issue as “a social question and beyond the scope” in [10] and provides some recommended values, like 0.1 or 0.01, and sometimes $\ln 2$, $\ln 3$. Some papers set the values of ϵ as from 0.5 to 1.5, or from 3 to 10. The authors in [20, 22] state that for an ϵ value, the probability of re-identification of an individual is not fixed, it is determined by data values in the given dataset and even by data for individuals outside the dataset.

Besides choosing the value of ϵ , the explanation of ϵ is also not an easy task. There is an on-going need to explain to non-specialists what differential privacy is, especially the meaning of privacy budget ϵ , and how much protection of the original data and utility of the private data will be for the various ϵ settings. This is very important to allow for the adoption of this technology in healthcare, especially when used for the protection of highly vulnerable patient groups [15].

Sensitivity Another important limitation of differential privacy regards the queries or functions’ sensitivities, which determine the scale of noise (perturbation) that should be injected to the original answers. So far, research works require that the queries or functions under differential privacy have low sensitivity that incurs small scale of perturbation, compared to the true answers [8]. Low sensitivity means that the outputs of queries or functions are not influenced severely by the change of any single record.⁴ One such concern of differential privacy is its application on numerical data, which may have high sensitivity and render the noisy output useless.

Another concern was mentioned by Muralidhar and Sarathy in [30, 34] and regards the inherent difficulty in computing queries’ sensitivity in unbounded domains. That raises two questions about not only the sensitivity calculation, but also about differential privacy verification as well. These questions may have been neglected in previous works, as they are limited to queries or functions with low or domain-independent sensitivities.

⁴Another definition is the removal or addition of a single data record.

3.7 Conclusion

In this chapter, we presented the basic concepts and mechanisms for achieving differential privacy. We surveyed state-of-the-art techniques in releasing private synthetic data and histograms for relational data from three categories: parametric algorithms, semi-parametric algorithms and non-parametric algorithms. Then, we investigated state-of-the-art methods for private data publication in the context of transaction/set-valued data and dynamic stream data. Last, we discussed a set of challenges pertaining to the offering of differentially-private data releases for health data, along with a set of promising directions for future research.

References

1. Ács, G., Castelluccia, C., Chen, R.: Differentially private histogram publishing through lossy compression. In: ICDM (2012)
2. Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar, K.: Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In: PODS (2007)
3. Bhaskar, R., Laxman, S., Smith, A., Thakurta, A.: Discovering frequent patterns in sensitive data. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 25–28 July 2010, pp. 503–512 (2010). doi:[10.1145/1835804.1835869](https://doi.acm.org/10.1145/1835804.1835869), <http://doi.acm.org/10.1145/1835804.1835869>
4. Chan, T.H., Shi, E., Song, D.: Private and continual release of statistics. ACM Trans. Inf. Syst. Secur. **14**(3), 26 (2011). doi:[10.1145/2043621.2043626](https://doi.acm.org/10.1145/2043621.2043626), <http://doi.acm.org/10.1145/2043621.2043626>
5. Chen, R., Mohammed, N., Fung, B.C.M., Desai, B.C., Xiong, L.: Publishing set-valued data via differential privacy. Proc.VLDB **4**(11), 1087–1098 (2011). <http://www.vldb.org/pvldb/vol4/p1087-chen.pdf>
6. Cormode, G., Procopiuc, C.M., Srivastava, D., Shen, E., Yu, T.: Differentially private spatial decompositions. In: ICDE (2012)
7. Cormode, G., Procopiuc, C.M., Srivastava, D., Tran, T.T.L.: Differentially private summaries for sparse data. In: ICDT, pp. 299–311 (2012)
8. Dankar, F.K., Emam, K.E.: The application of differential privacy to health data. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops, Berlin, Germany, 30 March 2012, pp. 158–166 (2012). doi:[10.1145/2320765.2320816](https://doi.acm.org/10.1145/2320765.2320816), <http://doi.acm.org/10.1145/2320765.2320816>
9. Dwork, C.: Differential privacy. In: Automata, Languages and Programming, Pt 2, vol. 4052. Springer, Berlin (2006)
10. Dwork, C.: Differential privacy: a survey of results. In: Proceedings of the 5th International Conference on Theory and Applications of Models of Computation, pp. 1–19 (2008)
11. Dwork, C.: Differential privacy in new settings. In: Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17–19, 2010, pp. 174–183 (2010). doi:[10.1137/1.9781611973075.16](https://doi.org/10.1137/1.9781611973075.16), <http://dx.doi.org/10.1137/1.9781611973075.16>
12. Dwork, C.: Differential privacy. In: Encyclopedia of Cryptography and Security, 2nd edn., pp. 338–340. Springer, Heidelberg (2011)
13. Dwork, C., Naor, M., Pitassi, T., Rothblum, G.N.: Differential privacy under continual observation. In: STOC, pp. 715–724 (2010)
14. Dwork, C., Mcsherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. Theory of Cryptography, New York, NY, USA, pp. 1–20 (2006)

15. Emam, K.E.: Guidelines for the deidentification of health information CRC Press; 1 edition (May 6, 2013)
16. Fan, L., Xiong, L., Sunderam, V.S.: Fast: differentially private real-time aggregate monitor with filtering and adaptive sampling. In: SIGMOD Conference, pp. 1065–1068 (2013)
17. Geng, Q., Viswanath, P.: The optimal mechanism in differential privacy. In: IEEE Symposium on Information Theory (2014)
18. Ghosh, A., Roughgarden, T., Sundararajan, M.: Universally utility-maximizing privacy mechanisms. In: Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, pp. 351–360 (2009)
19. Inan, A., Kantarcioğlu, M., Ghinita, G., Bertino, E.: Private record matching using differential privacy. In: EDBT (2010)
20. Kasiviswanathan, S.P., Rudelson, M., Smith, A., Ullman, J.: The price of privately releasing contingency tables and the spectra of random matrices with correlated rows. In: Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5–8 June 2010, pp. 775–784 (2010)
21. Kellaris, G., Papadopoulos, S., Xiao, X., Papadias, D.: Differentially private event sequences over infinite streams. In: Proceedings of the VLDB Endowment, PVLDB, vol. 7(12), pp. 1155–1166 (2014). <http://www.vldb.org/pvldb/vol7/p1155-kellaris.pdf>
22. Lee, J., Clifton, C.: Differential identifiability. In: The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, 12–16 August 2012, pp. 1041–1049 (2012)
23. Li, N., Qardaji, W.H., Su, D., Cao, J.: Prvbasis: frequent itemset mining with differential privacy. Proc. VLDB **5**(11), 1340–1351 (2012). http://vldb.org/pvldb/vol5/p1340_ninghuili_vldb2012.pdf
24. Li, H., Xiong, L., Jiang, X.: Differentially private synthesization of multi-dimensional data using copula functions. In: EDBT, pp. 475–486 (2014)
25. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: L -diversity: Privacy beyond k -anonymity. TKDD **1**(1), (2007). doi:[10.1145/1217299.1217302](https://doi.acm.org/10.1145/1217299.1217302), <http://doi.acm.org/10.1145/1217299.1217302>
26. Machanavajjhala, A., Kifer, D., Abowd, J.M., Gehrke, J., Vilhuber, L.: Privacy: theory meets practice on the map. In: Proceedings of the 24th International Conference on Data Engineering, ICDE, pp. 277–286 (2008)
27. McSherry, F.: Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: SIGMOD '09: Proceedings of the 35th SIGMOD International Conference on Management of Data, pp. 19–30. ACM, New York, NY, USA (2009). doi:<http://doi.acm.org/10.1145/1559845.1559850>
28. McSherry, F., Talwar, K.: Mechanism design via differential privacy. In: FOCS, pp. 94–103 (2007)
29. Mohammed, N., Chen, R., Fung, B.C.M., Yu, P.S.: Differentially private data release for data mining. In: KDD, pp. 493–501 (2011)
30. Muralidhar, K., Sarathy, R.: Does differential privacy protect terry gross privacy? In: Privacy in Statistical Databases. Springer, Berlin (2011)
31. Nissim, K., Raskhodnikova, S., Smith, A.: Smooth sensitivity and sampling in private data analysis. In: STOC (2007)
32. Privacy, H.: New guidance on de-identification methods under the HIPAA privacy rule. In: TMA Privacy Office Information Paper (2013)
33. Rastogi, V., Nath, S.: Differentially private aggregation of distributed time-series with transformation and encryption. In: SIGMOD Conference, pp. 735–746 (2010)
34. Sarathy R., Muralidhar, K.: Evaluating laplace noise addition to satisfy differential privacy for numeric data. Trans. Data Priv. **4**(1), 1–17 (2011)
35. Xiao, X., Wang, G., Gehrke, J.: Differential privacy via wavelet transforms. In: ICDE, pp. 225–236 (2010)
36. Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G.: Differentially private histogram publication. In: ICDE, pp. 32–43 (2012)

Chapter 4

Evaluating the Utility of Differential Privacy: A Use Case Study of a Behavioral Science Dataset

Raquel Hill

Abstract Healthcare datasets may contain information that participants and data collectors have a vested interest in keeping private. Additionally, social scientists who collect large amounts of medical data value the privacy of their survey participants. As they follow participants through longitudinal studies, they develop unique profiles of these individuals. A growing challenge for these researchers is to maintain the privacy of their study participants, while sharing their data to facilitate research. This chapter evaluates the utility of a differentially private dataset. There has been extensive work, and heightened public and governmental focus on the privacy of medical datasets. However, additional efforts are needed to help researchers and practitioners better understand the fundamental notion of privacy with regards to more recent techniques, like differential privacy. The results of the study align with the theory of differential privacy, showing that dimensionality is a challenge, and that when the number of records in the database is sufficiently larger than the number of cells covered by a database query, the number of statistical tests with results close to those performed on original data, increases.

4.1 Introduction

The sharing of medical data has many possible benefits, which include the creation of unified data visualizations for clinicians, the development of predictive and diagnostic support systems, reductions in institutional costs, and improvements in medical care. Medical data are often not shared with external parties, because even when transformations that are intended to preserve individual privacy are applied to raw data, e.g., when data is de-identified per the US HIPAA Safe Harbor rules or when data is transformed per the UK ICO's Anonymization Guidelines, sharing of such data may still introduce important privacy risks. This is because privacy does not only depend on the transformed data but also on context-specific information,

R. Hill (✉)

School of Informatics and Computing, Indiana University, Bloomington, IN, USA

e-mail: rhill@indiana.edu

Reason for Appointment
1.MDVIP PE
History of Present Illness
HPI Notes:
1) LIPID management, On Lovastatin 15 y, p Non STEMI in mid 90s w normal LHC; h/o arrhythmia > 20y; Vegan 5 y 2) Lifelong nasal congestion has lessened; now w decr nasal acuity and chronic prod post nasal cough 3) Sleep disorder, Scammel rec Provigil George declines and we review the medicine in detail; the narcolepsy does not impact his professional or private life hence no need for it; he is not driving anymore 4) recurrent fungus in inguinal area 5) post running headaches and neck pain.

Fig. 4.1 Excerpt from doctor's notes

such as shared demographic data of subjects, presence of fields that may be linked across other existing databases, social relationships among subjects, and profiles of the data recipient.

Preserving the privacy of medical data is even more challenging because the combination of unstructured textual narratives, video, images, audio and other health-focused data modalities, create unique medical profiles. Figure 4.1 contains an example of unstructured, narrative text from a patient's examination. It illustrates that these narratives may contain information about prescribed medications, medical disorders, lifestyle, and even parts of the patient's name. Prior research shows that self-reported medical diagnoses and prescription drug information makes records within a survey dataset unique [29]. Other uniqueness results show that simple demographics, including birthdate, gender and zipcode can be used in combination to identify 63–87 % of the U.S. population [15, 30]. Most medical information includes simple demographics. In addition, unique combinations of information have been used in numerous re-identification studies [25]. In principle, *any information that can be used to distinguish one person from another can be used for re-identifying data* [26]. Thus, unique combinations present a major challenge to any system that seeks to enforce a re-identification risk mitigation privacy policy.

Any information that distinguishes an individual or a record within a database may be used to identify the person to whom the information belongs. Even when traditional identifiers are removed, *the uniqueness of the resulting records may make a privacy breach easy to orchestrate or implement*. Eliminating all unique characterizations from the data may destroy the utility or usefulness of data. Therefore sharing these data with external parties may become a lengthy process of negotiating specific use agreements. Sharing of the data among researchers within the organization where the data resides also creates privacy risks. Therefore, when designing mechanisms to enforce privacy policies, one must consider the *type of policy*, the *data* and the *intended use for the data*.

Given the challenges of creating a privacy preserving medical dataset for publishing, this chapter focuses on structured data and the exploration of differential privacy (DP) mechanisms as a means for providing *privacy preserving mediated access* to this data. Specifically, instead of sharing a sanitized version of a dataset, the data owner provides an interface that enables individuals to query the data and receive a sanitized response. DP is a data perturbation technique that provides strong and formal privacy guarantees [8]. The essential goal of DP is to prevent a possible adversary from discovering whether or not some specific individual's data is present in a differentially private dataset, given some risk threshold. A statistical aggregation or computation is differentially private, or satisfies differential privacy, if the outcome is formally indistinguishable when run with and without any (one) particular record in the dataset. The level of indistinguishability is quantified as a privacy parameter. A common differential privacy mechanism is to add calibrated noise to a statistical aggregation or computation result, determined by the privacy parameter and the sensitivity of the computation to the inclusion or exclusion of any individual record. While traditional de-identification methods [13] or perturbation methods that add noise to individual values of data records [1] are subject to re-identification or data reconstruction attacks, depending on the background knowledge of an adversary, a differential privacy mechanism adds noise to statistical aggregation or computation outputs, and provides a strong and provable privacy guarantee with little assumptions on the background knowledge of an adversary.

Hill et al. [18] present a use case analysis framework for evaluating the utility of a differentially private dataset that contains self reported medical information. The data set resulted from a survey that was administered by behavioral scientists. The use case focused on evaluating the results of logistic regression by measuring the distance between odds ratios that were derived from the original and differentially-private data sets. This chapter extends that work and introduces the p-score, which captures whether a variable level has changed from significant to non-significant (or vice versa). The odds ratio (OR) distance and p-score are intended to mirror the judgement process of a behavioral science researcher using multivariate logistic regression to determine predictor utility. The OR distance is important because it tells the researcher how strongly the predictor level influences the response variable. The p-score is also important because non-significant predictor levels are often ignored when interpreting results. If a researcher only has access to the differentially private histogram, a non-zero value for either metric could result in very different conclusions about the data.

The remainder of the chapter is organized as follows. Section 4.2 provides the necessary background on differential privacy and other approaches for anonymizing data. Section 4.3 describes the use case, and the utility metrics that were used. The results are presented in Sect. 4.4. Section 4.5 extends the discussion of the results, and Sect. 4.6 concludes the chapter.

4.2 Background

This section provides an overview and background of privacy-preserving data publishing and privacy-preserving data mining techniques, with specific focus on k -anonymity [27, 28, 31] and differential privacy [7, 8], respectively.

Syntactic models, like k -anonymity, provide tools to publish useful information, while preserving data privacy. Syntactic models typically generalize a database until the data conforms to some set of rules, the syntactic condition. For example, given the quasi-identifier¹ attributes, k -anonymity requires that each of the released records be indistinguishable from $k - 1$ other records. k -anonymity assumes that quasi-identifiers are known at the time of data sanitization.

Privacy-preserving data mining techniques, like differential privacy, provide frameworks for querying databases in a privacy-preserving manner without releasing the actual database. Differential privacy is a perturbation technique that adds noise to the answers of queries on the data. It provides strong and formal privacy guarantees. A fundamental goal of differential privacy is to prevent a possible adversary from discovering whether or not some specific individual's data is present in a differentially private dataset, given some risk threshold.

The remainder of this section provides a high-level overview of k -anonymity, differential privacy, and a brief discussion of privacy applications.

4.2.1 Syntactic Models: k -Anonymity

Syntactic models, like k -anonymity, were created to reduce the risk of re-identification in published data. The risk of re-identification can be specified as the probability of selecting the matching record from the k possible records. When no additional information can be used to allow the attacker to exclude any of the k records, the risk of re-identification can be denoted as a simple probability, $1/k$. The example below is based on work by El Emam et al. [11]. In their work, they evaluate prescription drug data that had been provided by a pharmaceutical company. Table 4.1 provides a subset of the original data, while Table 4.2 contains the k -anonymized data. In this example, gender and year of birth are the assumed quasi-identifiers. The minimum equivalence class size is 2, which indicates that there is at least one other record that has the same quasi-identifier values, and the probability of re-identification is $1/2$.

The use of k -anonymity for more than two decades to publish data, including health data has resulted in guidelines, policies, court cases, and regulatory orders that define acceptable levels of risk [9]. Even in the presence of such vast practical use, the ever growing data landscape continues to challenge the limiting assumptions

¹A quasi-identifier is a feature or set of features that is sufficiently correlated with an entity and when combined with other such features create a unique identifier.

Table 4.1 Original prescription database (derived from [11])

Identifier	Quasi-identifier		Sensitive attribute
Name	Gender	Year of birth	DIN
Joe Brown	Male	1969	2046059
Jackson Brown	Male	1972	716839
Adam Smith	Male	1969	2241497
Amanda Smith	Female	1985	2046059
Mary Kinsey	Female	1976	392537
Timothy Janssen	Male	1985	363766
Scott Sanders	Male	1988	544981
Leslie Silverstone	Female	1977	293512
Carl Fredricks	Male	1969	544981
Cathy Nielson	Female	1985	596612
Eric Jefferson	Male	1977	725765

Table 4.2 Disclosed prescription database ($k = 2$) (derived from [11])

Quasi-identifier		Sensitive attribute
Gender	Decade of birth	DIN
Male	1960–69	2046059
Male	1970–79	716839
Male	1960–69	2241497
Female	1980–89	2046059
Female	1970–76	392537
Male	1980–89	363766
Male	1980–89	544981
Female	1970–79	293512
Male	1960–69	544981
Female	1980–89	596612
Male	1970–79	725765

of k -anonymity and its derivatives. For example k -anonymity assumes that all quasi-identifiers are known a-priori. Researchers argue that it is unrealistic to limit the adversary in this manner and that it may be possible to use any unique information to link an identity to a de-identified record [26].

For example, in Tables 4.1 and 4.2, the Drug Identification Number (DIN), although unique in some cases, is not considered a quasi-identifier. For this example, an attacker who knows the DIN for the drugs associated with an individual would be able to uniquely identify 64 % of the records. Knowing the DIN and Year-of-Birth would allow an attacker to re-identify the remaining 36 %.

Researchers have proposed a variety of approaches for managing biomedical data and protecting patient information. A thorough review of k -anonymity and its derivatives are presented in Gkoulalas-Divanis et al. [14]. El Emam et al. [10] consider extensions to k -anonymity in the context of two attack scenarios: one in which the attacker wants to re-identify a specific individual that he/she knows in the

anonymized dataset (called the *prosecutor scenario*), and one in which the attacker simply wants to demonstrate that an arbitrary individual from some population could be re-identified in the dataset (called the *journalist scenario*).

The best k -anonymity extension selects an appropriate k value using hypothesis testing and a truncated-at-zero Poisson distribution. While this method outperforms standard k -anonymity in terms of information loss (computed using the discernability metric) on their sample datasets, the authors acknowledge that increasing the number of quasi-identifiers may lead to unacceptable amounts of information loss, even for small values of k . Additionally, there was no discussion about how different values of the discernability metric may impact the results of common statistical analyses performed by consumers of an anonymized dataset (i.e., how much information loss will a logistic regression tolerate?).

4.2.2 Differential Privacy: Definition

Differential privacy (DP) has recently emerged as one of the strongest privacy guarantees for statistical data release. A statistical aggregation or computation is differentially private, or satisfies differential privacy, if the outcome is formally indistinguishable when run with and without any particular record in the dataset. The level of indistinguishability is quantified as a privacy parameter.

A common differential privacy mechanism is to add calibrated noise to a statistical aggregation or computation result determined by the privacy parameter and the sensitivity of the computation to the inclusion or exclusion of any individual record. Differential privacy provides a strong and provable privacy guarantee with little assumptions on the background knowledge of an adversary.

Differential privacy guarantees that if an adversary knows complete information of all the tuples in a dataset D except one, the output of a differentially private randomized algorithm should not give the adversary *too much additional information* about the remaining tuples. We say datasets D and D' differ in only one tuple, if we can obtain D' by removing or adding only one tuple from D . A formal definition of differential privacy is given as follows:

Definition 4.1 (ϵ -Differential Privacy [7]). Let \mathcal{A} be a randomized algorithm over two datasets D and D' differing in only one tuple, and let \mathcal{O} be any arbitrary set of possible outputs of \mathcal{A} . Algorithm \mathcal{A} satisfies ϵ -differential privacy if and only if the following holds:

$$\Pr[\mathcal{A}(D) \in \mathcal{O}] \leq e^\epsilon \Pr[\mathcal{A}(D') \in \mathcal{O}]$$

Intuitively, differential privacy ensures that the released output distribution of \mathcal{A} remains nearly the same whether or not an individual tuple is in the dataset.

The most common mechanism to achieve differential privacy is the Laplace mechanism [7] that adds a small amount of independent noise to the output of a

numeric function f to fulfil ϵ -differential privacy of releasing f , where the noise is drawn from a *Laplace distribution* with a probability density function $Pr[\eta = x] = \frac{1}{2b}e^{-\frac{|x|}{b}}$. Laplace noise has a variance $2b^2$ with a magnitude of b . The magnitude b of the noise depends on the concept of *sensitivity*, which is defined as follows:

Definition 4.2 (Sensitivity [7]). Let f denote a numeric function. The sensitivity of f is defined as the maximal L_1 -norm distance between the outputs of f over the two datasets D and D' , which differ in only one tuple. Formally,

$$\Delta_f = \max_{D,D'} \|f(D) - f(D')\|_1.$$

With the concept of sensitivity, the noise follows a zero-mean Laplace distribution with the magnitude $b = \Delta_f/\epsilon$. To fulfill ϵ -differential privacy for a numeric function f over D , it is sufficient to publish $f(D) + X$, where X is drawn from $\text{Lap}(\Delta_f/\epsilon)$.

For a sequence of differentially private mechanisms, the compositability [22] theorems (see below) guarantee the overall privacy.

Theorem 4.1 (Sequential Composition [22]). *For a sequence of n mechanisms M_1, \dots, M_n , where each mechanism M_i provides ϵ_i -differential privacy, the sequence of M_i provides $(\sum_{i=1}^n \epsilon_i)$ -differential privacy.*

Theorem 4.2 (Parallel Composition [22]). *If D_i are disjoint subsets of the original database and M_i provides α -differential privacy for each D_i , then the sequence of M_i provides α -differential privacy.*

Dankar et al. [6] provide a thorough treatment of the state-of-the-art in differential privacy. They also outline some of the limitations of the model and the various mechanisms that have been proposed to implement it. In addition, this work discusses several recent applications of differential privacy [12, 16, 19, 20, 23, 32, 33].

In this chapter, we evaluate the output of several DP processes, including cell-based [32], range query[33], and space partitioning[32]. The space partitioning approach differs from the basic cell-based and range query approaches in that it attempts to preserve the characteristics within the original data, by adding noise uniformly to cells that belong to a partitioned group.

Xiao et al.[32, 34] use a kd-tree (k -dimensional tree) to partition the data. A kd-tree is a space partitioning data structure for organizing data points in a k -dimensional space. First, the DP algorithm partitions the dataset D based on the domain and adds noise to each cell to create a synthetic dataset D' . D' is subsequently partitioned using a kd-tree algorithm. The resulting keys from the kd-tree partitioning are then used to subdivide the original dataset. Finally, Laplace noise is added to each partition's count. Each cell within a partition is assigned the value of its partition's $\text{noisy}_{\text{count}}/\beta$, where β is the number of cells within the partition. The perturbed dataset is used in the kd-tree phase of the algorithm, so as not to waste the privacy budget on accessing the original dataset multiple times during the partitioning phase.

Xiao et al. [32] evaluate the utility of their DP mechanism by comparing query counts of the original data to those of the differentially private data. We build upon this work by performing predictive analysis on a large social science dataset and the corresponding differentially private data. We derive our use cases from the actual analyses that were previously performed by the researchers who collected and evaluated the original data [17].

4.2.3 Applications

Various applications for securing access and reducing the privacy while sharing data have been proposed and implemented. Brown et al. [3] present a “distributed” query system designed to allow data holders to maintain physical control of their data. This system is contrasted with a centralized database, where users submit queries outside of their local firewalls (and also receive results from a remote server). Data privacy is maintained by physically co-locating the query system software with the data. The system does not attempt to anonymize data for access by a third party.

Murphy et al. [24] present the i2b2 system (integrating biology and the bedside), which provides graded access to patient data depending on the privacy level of the user. At the lowest level, only aggregate counts with Gaussian noise added is available. The problem of multiple queries allowing for convergence on the true count is discussed, but only solved for single user accounts (a user with multiple accounts could still discover the true count). For all other privacy levels, some form of anonymized patient data is available to the user, with the highest level having access to the original data. The de-identification methods for this data are not discussed (they are only listed as HIPAA compliant), and neither is the link between dataset dimensionality and re-identification.

Kushida et al. [21] perform a literature survey on de-identification and anonymization techniques. They focus on three main scenarios: free-text fields, images, and biological samples. For free-text fields, statistical learning-based systems provide the best performance (at or above manual de-identification). Anonymization techniques for images and biological samples are briefly discussed as well. Unfortunately, there is no mention of data privacy methods for datasets with coded fields (e.g., surveys) or dimensionality—i.e., how much easier is re-identification with high-dimensional data?

Bredfeldt et al. [2] develop a set of templates and a common zip-file directory structure for multi-site research collaborations with sensitive data. Additionally, some best practices are identified, such as not transferring the zip-file over e-mail or any unencrypted protocol. While the templates and structure provide support for collaboration, important details for protecting the data itself (e.g., via encryption or some privacy mechanism) are not discussed.

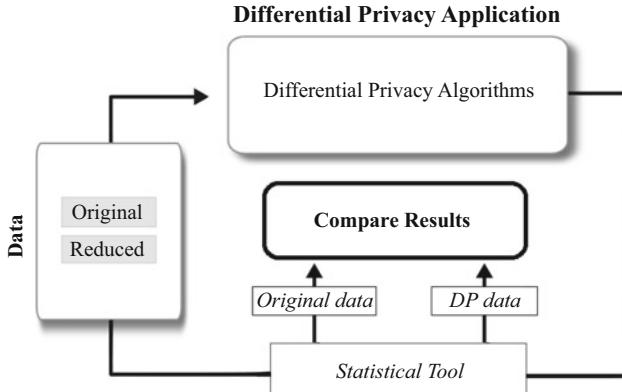


Fig. 4.2 Experiment flow chart

4.3 Methodology

Figure 4.2 illustrates a flow chart of the experimental setup. Hill et al. [18] also used this non-interactive querying approach. To evaluate differential privacy, a differentially private contingency table, a histogram of all possible attribute settings, is generated once. Data records are then reconstructed from the histogram bins. Statistical analysis is performed on both the original data and the DP data. The outcomes of the analyses are compared to assess the usefulness of the DP data.

The DP application implements several algorithms including two, *cell* and *basic*, that add Laplacian noise to each histogram cell independently, using the perturbation (privacy) ϵ parameter. The two algorithms differ in the way they compute the lambda parameter for Laplacian noise. In addition, the application implements a k-d tree algorithm that adds noise in two stages with the goal of preserving entropy [32, 34], and an algorithm that was designed to support range queries [33]. These algorithms are referred to as *k-d tree* and *adhoc* throughout the remainder of the text.

The parameter ϵ determines the level of privacy, with a lower value providing more privacy. For the k-d tree algorithm, the ET parameter determines the entropy, or uniformity, within each partition. The lower the value of ET, the less uniform partitions will be. In our experiments, we vary ϵ over the range [0.1, 2.0] and ET over the range [0.4, 1.0]. These ranges represent very low and very high privacy, and cover the true entropies of the datasets. For all our experiments, we fix the IG parameter at 0.0001.

Twenty five DP histograms are generated per algorithm, per each parameter setting, and the mean results from these histograms are compared to the original. Bonferroni correction is used to address the issue of multiple testing within our experiments. Our dataset contains 5887 participants who have answered 332 questions of a sexual health survey. The participants are a convenience sample of individuals within the USA. The survey contains multiple modules and most

modules are standardized sexual health *scales*—a questionnaire with which you hope to measure a psychological construct using multiple questions (i.e., items or indicators) [4].

The use cases are derived from work by Higgins et al. [17], and examine predictors of unprotected sex and unplanned pregnancy. Higgins et al. use logistic regression to evaluate the likelihood of a participant reporting an unplanned pregnancy and the likelihood of a participant reporting having had unsafe sex in the last 12 months (both binary outcomes). This chapter takes the same approach as Hill et al. [18] and evaluates both the full data model that contains all predictive variables that were used by Higgins et al. [17] and reduced data models that contain subsets of the predictive variables. The reduced models were used to demonstrate how dimensionality affects the usefulness of differentially private data. The full and reduced models are described below.

Predictor Variables

- **age:** Age of the participant (31 levels). Range is from 18–80 years old.
- **employ:** Employment status (4 levels): full-time, part-time, unemployed, temp/seasonal.
- **gender:** Gender (2 levels): male, female.
- **income:** Income level (4 levels): poor, lower, middle, upper.
- **relation:** Relationship status (3 levels): none, exclusive, non-exclusive.
- **sis5:** Use of safe sex products causes arousal loss (4 levels): strongly agree, agree, disagree, strongly disagree.
- **sis8:** Risk of pregnancy inhibits arousal (4 levels): strongly agree, agree, disagree, strongly disagree.

Response Variables

- **kisbq18:** Had unprotected vaginal sex in the last 12 months (binary).
- **kisbq20:** Ever had an unplanned pregnancy (binary).

Data Models

- **Full Set:** 7 predictor variables (age, gender, relation, employ, income, sis5, sis8) and 2 response variables (kisbq18, and kisbq20).
- **Reduced Set 1 (RS1):** 2 predictor variables (income, sis5) and 1 response variable (kisbq18).
- **Reduced Set 2 (RS2):** 2 predictor variables (relation, sis8) and 1 response variable (kisbq20).

4.3.1 Utility Measures

The size n of the database directly impacts the accuracy of a DP query. When $n \ll y$, where n is the number of records within the database and y is the number of cells covered by the query, the query results will be inaccurate [6, 8]. For the Full Set, RS1 and RS2, there are 190,464, 32, and 24 cells covered by the query, respectively. To evaluate the effect that the size of the database has on the accuracy of our results, we compare the results across the three datasets.

Data Distribution We compared the distributions of individual dimensions (or variables) for every DP histogram to the original distributions. For example, there were about 25 % more females than males in the original dataset. A utility-preserving DP histogram should have a similar male/female proportion (not necessarily the exact same number). All dimensions besides `age` were categorical, so we used a standard Pearson χ^2 goodness-of-fit test with the expected bin probabilities derived from the original data. The `age` variable was treated as continuous, so we performed a Kolmogorov-Smirnov (K-S) test instead of χ^2 . Effect sizes for dimensions besides `age` were computed with Cramér's V [5], and the K-S statistic itself was used as an effect size for variable `age`.

Logistic Regression Logistic regression produces an *odds ratio* (OR) and *p-value* for every *variable level* in the dataset (e.g., “full time” for the `employ`). All categorical variables had a *reference level* (shown in Tables 4.3 and 4.4). The `age` variable was treated as continuous, so it only had a single OR and p-value. For every set of parameters and variable level, we computed two metrics across all 25 runs: odds ratio distance and p-score.

The *odds ratio distance* is the absolute distance between the noisy and original odds ratios. For example, the odds ratio (M) for `employ` (full-time) in Table 4.3 was 1.56, meaning “full-time” employed males were more likely to have had unprotected sex in the last 12 months. If a noisy histogram had an OR of 0.56 for the same variable level and gender, the OR distance would be 1.0. This metric is larger when the noisy logistic results are farther away from the original dataset. To judge whether or not OR values are significantly different from the original, we used a one-sample Wilcoxon signed-rank test.

The *p-score* is 1 if the variable level has changed from *significant* to *non-significant* (or vice versa), and 0 otherwise. The `age` variable, for example, was highly significant for females in Table 4.4 ($p < 0.001$). If a noisy histogram had a p-value ≥ 0.05 for `age` (making it non-significant), then the p-score would be 1. Likewise, a change from non-significant to significant ($p < 0.05$) would result in a p-score of 1. A p-score of 0 would be given for a change from highly significant ($p < 0.001$) to significant ($p < 0.05$) or vice-versa. We summed all 25 p-scores for every variable level, and considered a p-score of 2 or more as significantly different from the original dataset. This allowed for 1 out of the 25 runs (4 %) to change significance and still be considered similar to the original variable level.

Table 4.3 Odds ratios (OR) and statistical significance (SS) for males (M) and females (F) in original data set (kisbq18)

Response	kisbq18 (unprotected sex)			
Predictor	OR (M)	SS (M)	OR (F)	SS (F)
age	1.12	$p < 0.001$ (***)	1.22	$p < 0.001$ (***)
employ (full time)	1.56		1.75	$p < 0.01$ (**)
employ (part time)	1.42		1.48	$p < 0.05$ (*)
employ (unemployed)	1.18		2.05	$p < 0.001$ (***)
employ (temporary)	Reference level			
income (poor)	0.85		1.29	
income (lower)	1.09		1.40	$p < 0.01$ (**)
income (middle)	1.06		1.29	$p < 0.05$ (*)
income (upper)	Reference level			
relation (exclusive)	1.53	$p < 0.01$ (**)	1.55	$p < 0.001$ (***)
relation (non-exclusive)	2.63	$p < 0.001$ (***)	2.18	$p < 0.01$ (**)
relation (none)	Reference level			
sis5 (strongly agree)	1.84	$p < 0.01$ (**)	3.23	$p < 0.001$ (***)
sis5 (agree)	1.13		2.56	$p < 0.001$ (***)
sis5 (disagree)	1.03		1.42	$p < 0.001$ (***)
sis5 (strongly disagree)	Reference level			

The number of asterisks indicates the strength of the results. Given three asterisks as opposed to two or one, it is less likely to incorrectly reject the null hypothesis

The two metrics above are intended to mirror the judgement process of a social science researcher using multivariate logistic regression to determine predictor utility. The OR distance is important because it tells the researcher how strongly the predictor level influences the response variable. The p-score is also important because non-significant predictor levels are often ignored when interpreting results. If a researcher only has access to the differentially-private histogram, a non-zero value for either metric could result in very different conclusions about the data.

4.4 Results

Before evaluating the results, we first compare the size of our database n to the number of cells that are covered for each use case. Recall that $n = 5887$ in the original dataset, while the cell coverage for the Full Set use case is 190,464 cells. Therefore, as we decrease the amount of noise that is added, we do not expect the odds ratio distances for the logistic regressions to change significantly.

When considering the reduced set use cases, RS1 and RS2, the number of records n is significantly larger than the cell coverage. Therefore, as we alter the ϵ parameter to decrease noise, we expect that more variable counts will be preserved, and a decrease in the distance between the original and DP odds ratios.

Table 4.4 Odds ratios (OR) and statistical significance (SS) for males (M) and females (F) in original data set (kisbq20)

Response	kisbq20 (unplanned pregnancy)			
Predictor	OR (M)	SS (M)	OR (F)	SS (F)
age	1.21	$p < 0.001 (***)$	1.52	$p < 0.001 (***)$
employ (full time)	1.31		1.65	$p < 0.05 (*)$
employ (part time)	0.80		1.24	
employ (unemployed)	1.02		1.64	$p < 0.05 (*)$
employ (temporary)	Reference level			
income (poor)	1.6		2.12	$p < 0.001 (***)$
income (lower)	1.28		1.55	$p < 0.001 (***)$
income (middle)	1.09		1.32	$p < 0.05 (*)$
income (upper)	Reference level			
relation (exclusive)	1.68	$p < 0.01 (**)$	1.91	$p < 0.001 (***)$
relation (non-exclusive)	2.56	$p < 0.001 (***)$	2.24	$p < 0.001 (***)$
relation (none)	Reference level			
sis8 (strongly agree)	Reference level			
sis8 (agree)	1.11		1.03	
sis8 (disagree)	1.72	$p < 0.01 (**)$	1.31	$p < 0.05 (*)$
sis8 (strongly disagree)	1.91	$p < 0.01 (**)$	1.41	$p < 0.05 (*)$

Overall, the two most important determinants of utility preservation were: (1) the dimensionality of the dataset (i.e., the number of variables), and (2) the entropy threshold of the k-d tree algorithm.

4.4.1 Variable Distributions

The differentially private distributions of individual dimensions (variables) are compared to those of the original data. If differentially private variable distributions differ significantly, it is unlikely that other kinds of utility (e.g., logistic regression) will be preserved. For all variables besides `age`, we judged a variable's distribution to be significantly different using Pearson's χ^2 goodness-of-fit test with bin probabilities derived from the original data. Effect size for these variables was computed with Cramér's V [5] using the corresponding χ^2 statistic. We treated the `age` variable as continuous, and therefore used a Kolmogorov-Smirnov (K-S) test with the K-S statistic serving as an effect size.

4.4.1.1 Full Set

For the full dataset, we found that *none*(!) of the noisy histograms contained a variable distribution that was similar to the original dataset. When looking at the

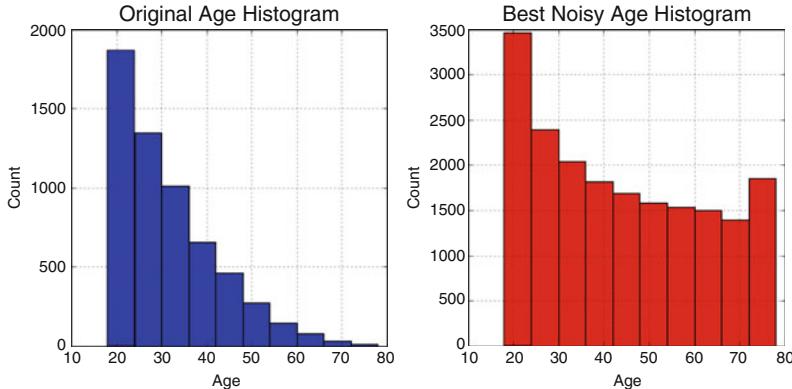


Fig. 4.3 Histogram of ages from original data (*left*) and using k-d tree algorithm with $\epsilon = 2.0$, ET = 0.677 (*right*)

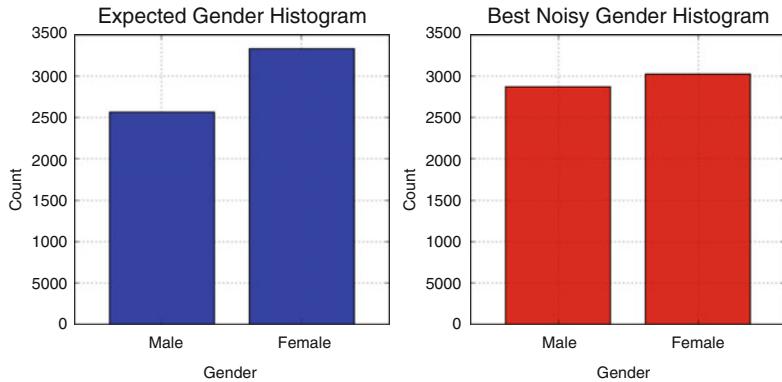


Fig. 4.4 Histogram of genders from original data (*left*) and using cell-based algorithm with $\epsilon = 2.0$ (*right*)

noisy histograms with the lowest effect size, it is immediately apparent why this is the case. Figure 4.3 shows the distribution of age from the original histogram and also from the DP histogram with the smallest effect size (K-S statistic). The DP histogram was generated using the k-d tree algorithm with ϵ set to 2.0 (low noise) and the entropy threshold (ET) set to the entropy of the original histogram. Visually, we can see that *important properties of the distribution have been lost*, such as the long tail. Figure 4.4 shows a similar story for the gender variable. Even though this histogram has the smallest effect size (Cramér's V), the noisy distribution is much more uniform than the original.

Effect size appears to be mostly driven by ϵ for all DP algorithms. While effect sizes decrease as ϵ increases for both age and other variables, they are still fairly large. Given the reduced set results, it appears that the dimensionality of the full dataset is really the biggest contributor to the large effect sizes and significantly different variable distributions.

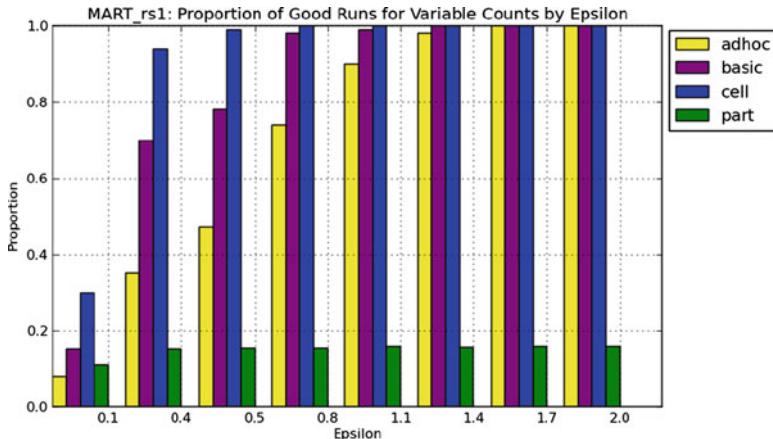


Fig. 4.5 Proportion of variable counts vs. ϵ for all algorithms (for the first reduced dataset)

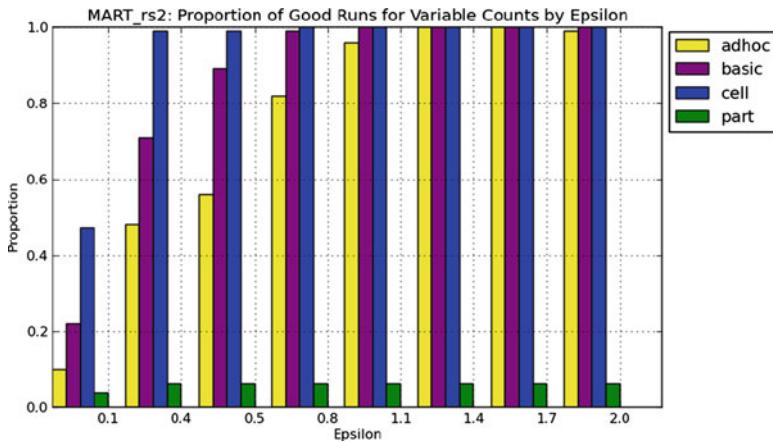


Fig. 4.6 Proportion of variable counts vs. ϵ for all algorithms (for the second reduced dataset)

4.4.1.2 Reduced Sets

Whereas all of the DP variable distributions from the full dataset were significantly different from their originals, *the DP variable distributions for the reduced datasets had far better utility*. For certain values of ϵ , 100 % of the variable distributions were preserved. Figures 4.5 and 4.6 compare the performance of the four algorithms for both reduced datasets.

Figures 4.7 and 4.8 depict the detailed performance of the k-d tree algorithm for MART_rs1 and MART_rs2. As the figures show, variable distributions are similar to the original when the entropy threshold is greater than or equal to that of the original data. As much as 75 % of all variables are preserved for MART_rs1 when $entropy_threshold = 1$ and $\epsilon > 0.1$.

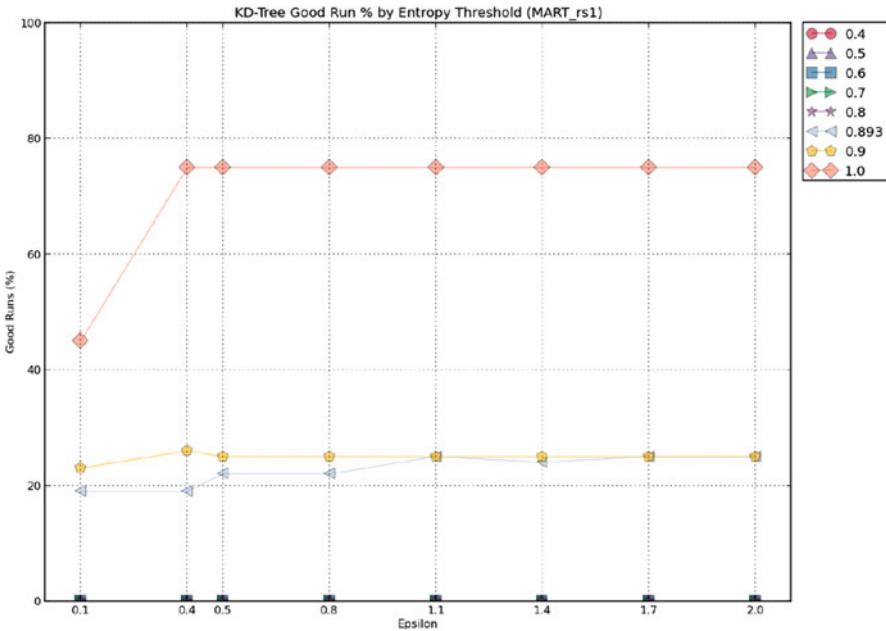


Fig. 4.7 Proportion of variable counts preserved vs. ϵ for k-d tree (for MART_rs1)

The ad-hoc, basic and cell-based methods had broader variable coverage with utility being preserved for all variables, while the kd-tree method had more similar runs for individual variables, such as kisbq18 and relation. Upon further investigation, we see that as ϵ increases, effect size decreases (Fig. 4.9). This holds true for all DP algorithms on both reduced datasets. The same is not true, however, for entropy threshold (ET). Instead, for kd-tree we find that all similar runs had their ETs at the original dataset entropy or above. So having a high ET was a prerequisite for a similar run, while ϵ merely modulated effect size.

4.4.2 Multivariate Logistic Regression

The second measure of data utility uses multivariate logistic regression. The use cases were derived from Higgins et al. [17]. For the full dataset, there were two response variables: kisbq18 (unsafe sex) and kisbq20 (unplanned pregnancy). Separate regressions were run for males and females, making a total of 4 use cases in the full dataset (2 response variables \times 2 genders). The reduced datasets each had a single response variable, so only 2 regressions were run per noisy histogram (male and female). Table 4.5 shows the predictors and response variables for each dataset.

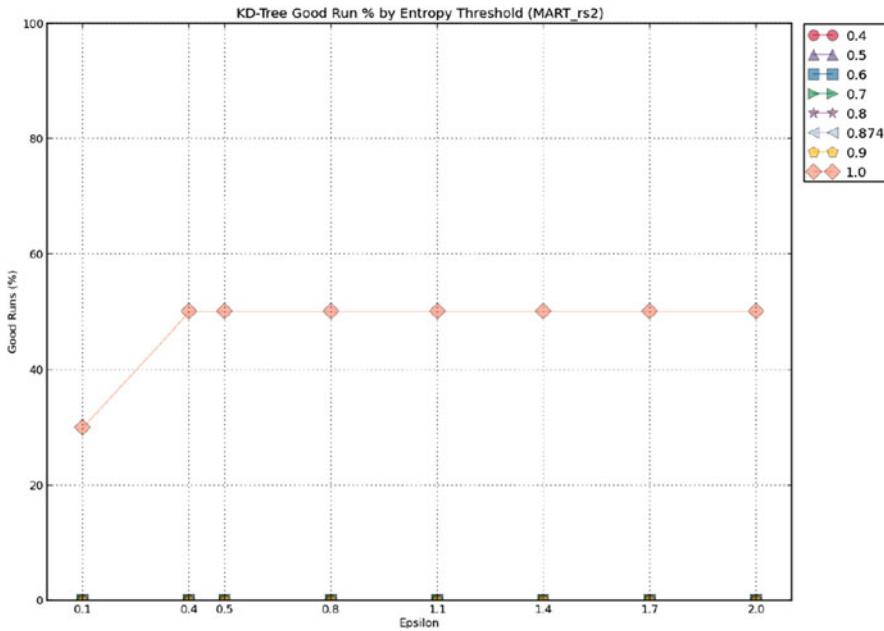


Fig. 4.8 Proportion of variable counts preserved vs. ϵ for k-d tree (for MART_rs2)

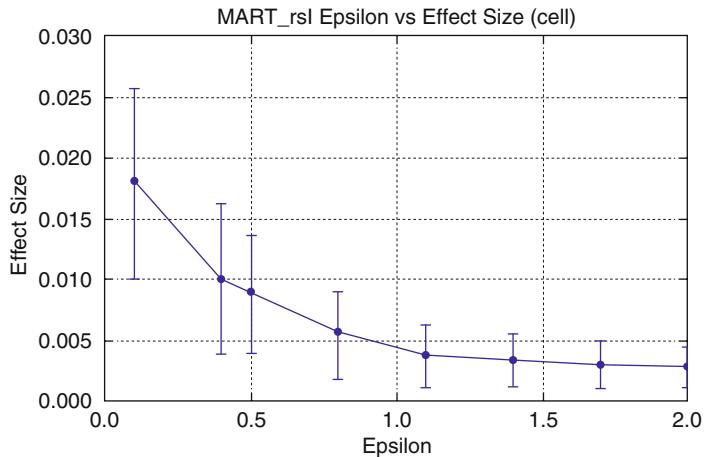


Fig. 4.9 Effect size versus ϵ for RS1 cell-based runs that were similar

For every set of parameters and variable level, the odds ratio distance and *p-score* were computed. The *p-score* is 1 if the variable level has changed from significant to non-significant (or vice versa), and 0 otherwise. The *p-score* is summed across the 25 noisy histograms for a parameter setting, and a $p\text{-score} < 2$ is considered a good run and utility preserving. This allowed for 1 out of the 25 runs (4 %) to change significance and still be considered similar to the original variable level.

Table 4.5 Logistic regressions for each dataset

Set	Query	Predictors	Resp.
Full	Unsafe	Age, employ, income, relation, sis5	kisbq18
Full	Preg	Age, employ, income, relation, sis8	kisbq20
RS1	Unsafe	Income, sis5	kisbq18
RS2	Preg	Relation, sis8	kisbq20

Queries are “unsafe” (unsafe sex) and “preg” (unplanned pregnancy)

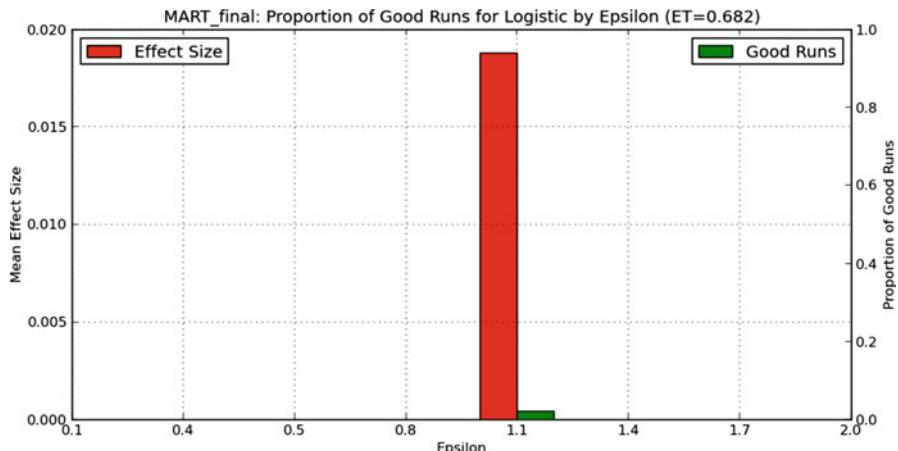


Fig. 4.10 Logistic results for MART_final for k-d tree, effect size and proportion of good runs versus the DP ϵ parameter

4.4.2.1 Noisy Results

Across all datasets and parameter settings, only 4.6 % of the variable levels were similar to the originals. Of those similar levels, 24.5 % are from the adhoc algorithm, 32.4 % from the basic, 41.4 % from the cell and 1.7 % from the kd-tree algorithm. Only the kd-tree algorithm provided utility preserving levels for the full dataset, MART_final. Figure 4.10 shows these results. The utility is preserved for SIS5, level = disagree. Interestingly, this is the most frequently provided response to the question, “Use of safe sex products causes arousal loss”. The utility is preserved when $\epsilon = 1.1$ and $entropy_threshold = 0.682$, which is the entropy of the original dataset. Also note that an effect size of 0.020 shows a strong relationship between the original odds ratio for this variable level and the mean of the odds ratio for the 25 samples when $\epsilon = 1.1$, and $entropy_threshold = 0.682$.

Most of the utility-preserving levels were from the reduced datasets (about 88 %). Figure 4.11 compares the proportion of good runs for each algorithm and value of ϵ for the MART_rs1 dataset. As the figure shows, the cell-based algorithm

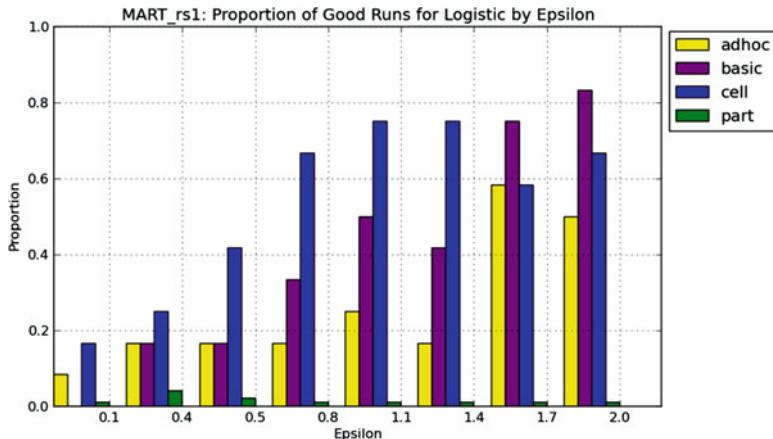


Fig. 4.11 Logistic results for MART_rs1 for k-d tree, proportion of good runs versus the DP ϵ parameter

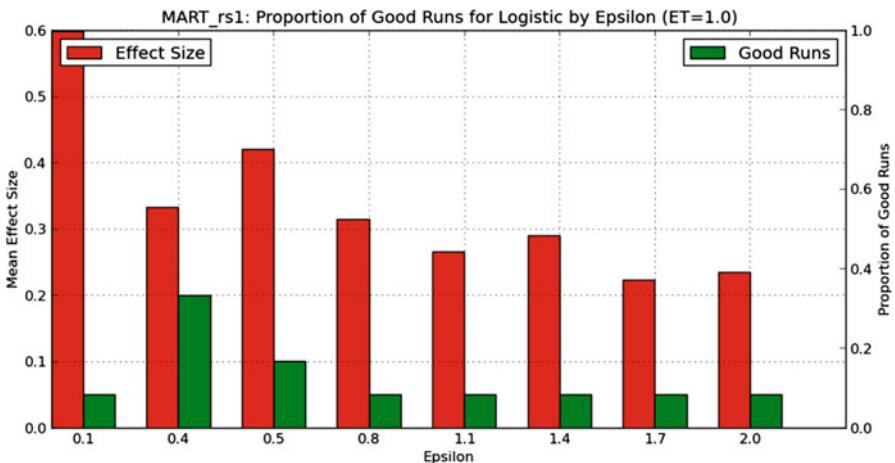


Fig. 4.12 Logistic results for MART_rs1 for k-d tree, effect size and proportion of good runs versus the DP ϵ parameter for entropy_threshold = 1.0

outperforms all other algorithms for ϵ values 0.1–1.4, which includes the higher noise values. The wavelet basic algorithm performs slightly better than the cell algorithm for ϵ values 1.7 and 2.0.

Figure 4.11 also shows that the kd-tree algorithm does not perform as well as the other algorithms and that the performance is somewhat consistent regardless of the value of ϵ . Figure 4.12 provides more detail and shows that all utility preserving levels for the k-d tree algorithm for MART_rs1 occur when the entropy_threshold = 1. The corresponding entropy_threshold for the original MART_rs1 dataset is 0.893. When $\epsilon=0.4$, 33 % of the runs preserve utility, which is a significant increase

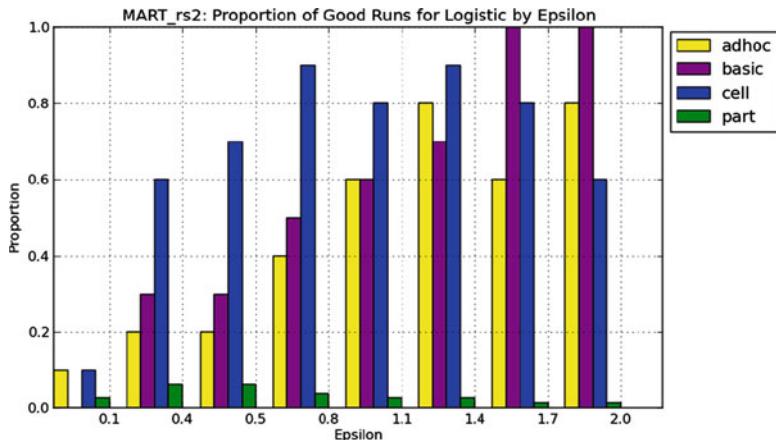


Fig. 4.13 Logistic results for MART_rs2, proportion of good runs versus the DP ϵ parameter

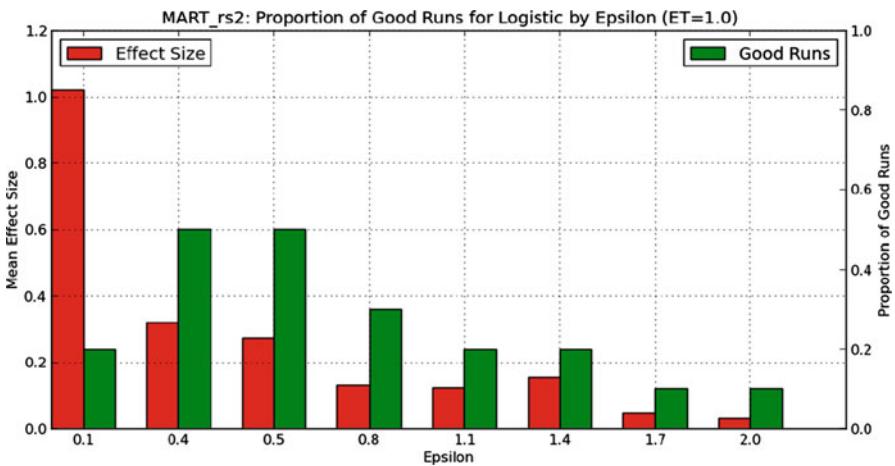


Fig. 4.14 Logistic results for MART_rs2 for k-d tree, effect size and proportion of good runs versus the DP ϵ parameter for entropy_threshold = 1.0

in comparison with the results from the full dataset, MART_final. Additionally, all utility preserving levels occur for variable SIS5 for levels agree and strongly agree, with twice as many occurring for the latter.

We report similar results for the second reduced dataset, MART_rs2. As Fig. 4.13 shows, the cell-based algorithm performs significantly better for six of the eight values of ϵ and has 90 % of its runs preserve utility for $\epsilon = 0.8$ and 1.4. The wavelet basic algorithm performs slightly better than cell for ϵ values 1.7 and 2.0 with 100 % of its runs preserving utility for these values.

Figure 4.13 also shows that the kd-tree algorithm does not perform as well as the other algorithms and that the performance is somewhat consistent regardless of the value of ϵ . Figure 4.14 provides more detail and shows that all utility-preserving levels for the k-d tree algorithm for MART_rs2 occur when $entropy_threshold = 1$. The corresponding entropy_threshold for the original MART_rs2 dataset is 0.874. When $\epsilon=0.4$ and 0.5, 50 % of the runs preserve utility, which is a significant increase in comparison with the results from the full dataset, MART_final. Additionally, the utility preserving levels occur for the relationship and SIS8 variables, with 52 % for exclusive, 38 % for nonexclusive and 10 % for SIS8 disagree.

For the reduced sets, the average effect size decreased as ϵ increased (ad-hoc, basic, cell-based), and as ϵ and entropy threshold increased (k-d tree). Effect size for the full dataset, however, was not strongly correlated with ϵ . Having an entropy threshold at or above the original histogram entropy was strong predictor of utility preservation across all datasets. As with the variable distribution utility measure, the reduced datasets *only* had utility-preserving variable levels with the k-d tree algorithm when the entropy threshold met this condition.

4.5 Discussion

The experiment hypothesis was that if the variable distributions were not preserved, then it would be unlikely that the results of other classifiers, like logistic regression would hold. The results show that none of the variable distributions for our full dataset, MART_final, were preserved. Consequently, the utility of the logistic results for only one variable was preserved for the full dataset. The corresponding noisy histogram was generated using the k-d tree partitioning algorithm, with $\epsilon = 1.1$ and $ET = 0.682$ (the entropy of the original dataset). These logistic results are similar in nature to those presented by Hill et al. [18], which show that for the full dataset, the odds ratio distances do not change even when the amount of noise added is reduced. As previously stated, the query coverage for the full dataset is significantly larger than the number of records. Under such conditions, the results show that dimensionality limits the creation of utility preserving DP datasets.

The overall utility preservation performance for the noisy reduced datasets was significantly stronger than the full dataset. We observed a strong performance for the basic, adhoc and cell-based algorithms for the variable utility distribution, with preservation rates approaching 100 % for higher values of ϵ (i.e., less noise). The k-d tree algorithm also produced more utility-preserving histograms when the entropy threshold parameter was set to a value that was greater than or equal to the entropy threshold of the original dataset. As much as 75 % of the variable distributions were preserved for MART_rs1 when using the k-d tree algorithm with $ET = 1$. The same trend was observed for the logistic, *p-score* utility measure for the reduced sets.

4.6 Conclusion

One challenge to assessing the utility of DP data is specifying what constitutes equivalent results. While results by Hill et al. [18] show that in some cases distance measures and error rates approach zero, this does not sufficiently articulate equivalence. When discussing these results with behavioral scientists, we found that this measure of utility may not translate. Therefore, this chapter extends prior work and incorporates the *p-score*, which seeks to model the decision making process of behavioral scientists, and bridge the gap between a distance measure of accuracy and the way in which behavioral scientists evaluate the quality of their results.

The results confirm that dimensionality is a major challenge for differentially private algorithms, especially when the number of records in the database is sufficiently less than the number of cells covered by the query. Given the two reduced sets, the query cell coverage for MART_rs1 is greater than for MART_rs2, indicating that some variables in MART_rs1 have a larger range than those in MART_rs2. This larger range provides more opportunities for cells to have zero counts. When the DP application adds noise to these cells, the resulting distribution for this dimension is altered. Therefore, when the range is large, the resulting utility of the differentially private data is less.

Acknowledgements This work is funded by NSF grants CNS-1012081.

References

1. Aggarwal, C.C.I., Yu, P.S.: A survey of randomization methods for privacy-preserving data mining. In: Privacy-Preserving Data Mining. Advances in Database Systems, vol. 34, pp. 137–156. Springer, New York (2008)
2. Bredfeldt, C.E., Butani, A.L., Pardee, R., Hitz, P., Padmanabhan, S., Saylor, G.: Managing personal health information in distributed research environments. BMC Med. Inform. Decis. Mak. **13**, 116 (2013). <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3851487/>
3. Brown, J., Holmes, J., Shah, K., Hall, K., R., L., Platt, R.: Distributed health data networks. Med. Care **48**(6 Suppl), S45–S51 (2010)
4. Clark, L., Watson, D.: Constructing validity: basic issues in objective scale development. Psychol. Assess. **7**(3), 309–319 (1995)
5. Cramér, H.: Mathematical Methods of Statistics, vol. 9. Princeton University Press, Princeton (1945)
6. Dankar, F.K., El Emam, K.: The application of differential privacy to health data. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops, pp. 158–166. ACM (2012)
7. Dwork, C.: Differential privacy. In: Automata, Languages and Programming, pp. 1–12. Springer, Berlin (2006)
8. Dwork, C.: Differential privacy: a survey of results. In: Theory and Applications of Models of Computation, pp. 1–19. Springer, Berlin (2008)
9. El Emam, K., Arbuckle, L.: Anonymizing Health Data, 1st edn. O'Reilly Media, Sebastopol, CA, USA (2013)
10. El Emam, K., Dankar, F.: Protecting privacy using k-anonymity. J. Am. Med. Inform. Assoc. **15**(5), 627–637 (2008)

11. El Emam, K., Dankar, F., Vaillancourt, R., Roffey, T., Lysyk, M.: Evaluating the risk of re-identification of patients from hospital prescription records. *Can. J. Hosp. Pharm.* **62**(4), 307–319 (2009)
12. Feldman, D., Fiat, A., Kaplan, H., Nissim, K.: Private coresets. In: Proceedings of the 41st Annual ACM Symposium on Theory of Computing, pp. 361–370. ACM (2009)
13. Fung, B., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: a survey of recent developments. *ACM Comput. Surv. (CSUR)* **42**(4), 14 (2010)
14. Gkoulalas-Divanis, A., Loukides, G., Sun, J.: Publishing data from electronic health records while preserving privacy: a survey of algorithms. *J. Biomed. Inform.* **50**, 4–19 (2014)
15. Golle, P.: Revisiting the uniqueness of simple demographics in the US population. In: Proceedings of the 5th ACM Workshop on Privacy in Electronic Society, WPES '06, pp. 77–80. ACM, New York (2006). doi:[10.1145/1179601.1179615](https://doi.acm.org/10.1145/1179601.1179615). <http://doi.acm.org/10.1145/1179601.1179615>
16. Götz, M., Machanavajjhala, A., Wang, G., Xiao, X., Gehrke, J.: Privacy in search logs (2009, preprint). arXiv:0904.0682
17. Higgins, J.A., Tanner, A.E., Janssen, E.: Arousal loss related to safer sex and risk of pregnancy: implications for women's and men's sexual health. *Perspect. Sex. Reprod. Health* **41**(3), 150–157 (2009)
18. Hill, R., Hansen, M., Janssen, E., Sanders, S.A., Heiman, J.R., Xiong, L.: A quantitative approach for evaluating the utility of a differentially private behavioral science dataset. In: Proceedings of the IEEE International Conference on Healthcare Informatics. IEEE (2014)
19. Inan, A., Kantarcioglu, M., Ghinita, G., Bertino, E.: Private record matching using differential privacy. In: Proceedings of the 13th International Conference on Extending Database Technology, pp. 123–134. ACM (2010)
20. Korolova, A., Kenthapadi, K., Mishra, N., Ntoulas, A.: Releasing search queries and clicks privately. In: Proceedings of the 18th International Conference on World wide Web, pp. 171–180. ACM (2009)
21. Kushida, C.A., Nichols D.A., Jadrnicke, R., Miller, R., Walsh, J.K., Griffin, K.: Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med. Care* **50**, S82–S101 (2012)
22. McSherry, F.D.: Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, SIGMOD '09, pp. 19–30. ACM, New York (2009). doi:[10.1145/1559845.1559850](https://doi.acm.org/10.1145/1559845.1559850). <http://doi.acm.org/10.1145/1559845.1559850>
23. McSherry, F., Mironov, I.: Differentially private recommender systems: building privacy into the net. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 627–636. ACM (2009)
24. Murphy, S.N., Gainer, V., Mendis, M., Churchill, S., Kohane, I.: Strategies for maintaining patient privacy in i2b2. *J. Am. Med. Inform. Assoc.* **13**(Suppl), 103–108 (2011)
25. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: Proceedings of the 2008 IEEE Symposium on Security and Privacy, SP '08, pp. 111–125. IEEE Computer Society, Washington, DC (2008). doi:[10.1109/SP.2008.33](https://doi.org/10.1109/SP.2008.33). <http://dx.doi.org/10.1109/SP.2008.33>
26. Narayanan, A., Shmatikov, V.: Myths and fallacies of personally identifiable information. *Commun. ACM* **53**(6), 24–26 (2010). doi:[10.1145/1743546.1743558](https://doi.acm.org/10.1145/1743546.1743558). <http://doi.acm.org/10.1145/1743546.1743558>
27. Samarati, P.: Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.* **13**(6), 1010–1027 (2001). doi:[10.1109/69.971193](https://doi.org/10.1109/69.971193). <http://dx.doi.org/10.1109/69.971193>
28. Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information (abstract). In: Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, PODS '98, p. 188. ACM, New York (1998). doi:[10.1145/275487.275508](https://doi.acm.org/10.1145/275487.275508). <http://doi.acm.org/10.1145/275487.275508>

29. Solomon, A., Hill, R., Janssen, E., Sanders, S.A., Heiman, J.R.: Uniqueness and how it impacts privacy in health-related social science datasets. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, pp. 523–532. ACM (2012)
30. Sweeney, L.: Uniqueness of simple demographics in the U.S. population. In: Technical Report: LIDAP WP4, Carnegie Mellon (2000)
31. Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Fuzziness Knowl. Based Syst.* **10**(5), 557–570 (2002)
32. Xiao, Y., Xiong, L., Yuan, C.: Differentially private data release through multidimensional partitioning. In: Secure Data Management, pp. 150–168. Springer, Berlin (2010)
33. Xiao, X., Wang, G., Gehrke, J.: Differential privacy via wavelet transforms. *IEEE Trans. Knowl. Data Eng.* **23**(8), 1200–1214 (2011)
34. Xiao, Y., Xiong, L., Fan, L., Goryczka, S., Li, H.: DPcube: differentially private histogram release through multidimensional partitioning. *Transactions on Data Privacy* **7**(3), 195–222 (2014)

Chapter 5

SECRETA: A Tool for Anonymizing Relational, Transaction and RT-Datasets

Giorgos Poulis, Aris Gkoulalas-Divanis, Grigoris Loukides,
Spiros Skiadopoulos, and Christos Tryfonopoulos

Abstract Publishing medical datasets about individuals, in a privacy-preserving way, has led to a significant body of research. Meanwhile, algorithms for anonymizing such datasets, with relational or set-valued (a.k.a. transaction) attributes, in a way that preserves *data truthfulness*, are crucial to medical research. Selecting, however, the most appropriate algorithm is still far from trivial, and tools that assist data publishers in this task are needed. To highlight this need, we initially provide a brief description of the popular anonymization algorithms and the most commonly used metrics to quantify data utility. Next, we present a system that we have designed, which is capable of applying a set of anonymization algorithms, enabling data holders, including medical researchers and healthcare organizations, to test the effectiveness and efficiency of different methods. Our system, called SECRETA, allows evaluating a specific anonymization algorithm, comparing multiple anonymization algorithms, and combining algorithms for anonymizing datasets with both relational and transaction attributes. The analysis of the algorithms is performed in an interactive and progressive way, and results, including attribute statistics and various data utility indicators, are summarized and presented graphically.

G. Poulis (✉)

NOKIA Research and Development, Athens Technology Center, Athens, Greece
e-mail: poulis@uop.gr

A. Gkoulalas-Divanis

Smarter Cities Technology Centre, IBM Research, Ireland
e-mail: arisdiva@ie.ibm.com

G. Loukides

School of Computer Science and Informatics, Cardiff University, Cardiff, UK
e-mail: g.loukides@cs.cf.ac.uk

S. Skiadopoulos • C. Tryfonopoulos

Department of Informatics and Telecommunications, University of Peloponnese,
Tripoli, Greece
e-mail: spyros@uop.gr; trifon@uop.gr

5.1 Introduction

Several medical studies require datasets which contain not only set-valued information (such as diagnosis codes of patients), but also relational information (such as patient demographics). This kind of datasets, henceforth referred to as *RT-datasets*, cannot be published in their original form, as it is provably easy for attackers to link a record to the corresponding patient [14]. Consider the original dataset of Table 5.1. The ids and names of the patients will not be disclosed in the published counterpart of the dataset; they are only illustrated for the ease of presentation. If an attacker knows that John is from France and has been diagnosed with Flu and Herpes, it is easy to link John to the first patient of Table 5.1. This type of attack is called *identity disclosure*, and can be issued either on the relational part of the dataset, or on the transaction part of the data, *or in both*.

The relational attributes in a dataset that are not direct identifiers, but can still be used (in combination) by an attacker to uniquely identify a user in the dataset, are called *quasi-identifiers* (QI). For example, in Table 5.1, we consider Age, Origin and Gender to be the quasi-identifiers. In order to protect a dataset from identity disclosure, the dataset has to be anonymized prior to data publishing.

Several methods (e.g., [4, 8, 11, 22]) have been proposed for performing data anonymization, by preventing the disclosure of individuals' private and sensitive information, while maintaining *data truthfulness*. Different to noise addition approaches, such as differential privacy [3], these methods lead to anonymized datasets that can be accurately analysed at a record level, without falsifying the data. For instance, Table 5.2c illustrates the anonymization of Table 5.1 using the method proposed by Poulis et al. [14]. In this dataset, an attacker cannot uniquely identify John, since his record can be either Id 1 or 2.

Let us consider a medical researcher who needs to conduct three medical studies. The first study is based on the demographics of patients, the second on the diagnosis codes of patients, and the third examines the correlation between certain diagnoses and patients' demographics. A data owner (e.g., a healthcare clinic) cannot disclose the data to the researcher in its original form, as the patients' privacy would be compromised. Also, the data owner does not typically have the necessary technical expertise to apply anonymization algorithms on the data.

Table 5.1 An example of an RT-dataset containing patient demographics and diseases

Id	Name	Age	Origin	Gender	Disease
1	John	19	France	Male	Asthma Flu Herpes
2	Steve	22	Greece	Male	Asthma Flu Eczema
3	Mary	28	Germany	Female	Flu Herpes
4	Zoe	39	Spain	Female	Asthma Flu Eczema
5	Ann	70	Algeria	Female	Asthma
6	Jim	55	Nigeria	Male	Asthma Flu

Table 5.2 (a) A 2-anonymous dataset with respect to relational attributes, (b) a 2^2 -anonymous dataset with respect to the transaction attribute, and (c) a $(2, 2^2)$ -anonymous dataset

Id	Age	Origin	Gender
1	[19:22]	Europe	Male
2	[19:22]	Europe	Male
3	[28:39]	Europe	Female
4	[28:39]	Europe	Female
5	[55:70]	Africa	All
6	[55:70]	Africa	All

(a)

Id	Disease
1	Asthma Flu (Herpes, Eczema)
2	Asthma Flu (Herpes, Eczema)
3	Flu (Herpes, Eczema)
4	Asthma Flu (Herpes, Eczema)
5	Asthma
6	Asthma Flu

(b)

Id	Age	Origin	Gender	Disease
1	[19:22]	Europe	Male	Asthma Flu (Herpes, Eczema)
2	[19:22]	Europe	Male	Asthma Flu (Herpes, Eczema)
3	[28:39]	Europe	Female	(Asthma, Flu) (Herpes, Eczema)
4	[28:39]	Europe	Female	(Asthma, Flu) (Herpes, Eczema)
5	[55:70]	Africa	All	(Asthma, Flu)
6	[55:70]	Africa	All	(Asthma, Flu)

(l)

In this chapter, we provide an overview of SECRETA [17], a system that allows the healthcare clinic to release an anonymized version of the data to the researcher. SECRETA supports datasets holding either relational or transaction data, as well as RT-datasets. Moreover, it supports a variety of anonymization algorithms and the most common utility metrics. In more detail, Table 5.2a presents an anonymous counterpart of the dataset shown in Table 5.1 with respect to the demographics of the patients, while Table 5.2b presents an anonymous dataset with respect to their diagnosis codes. Last, Table 5.2c presents an anonymous version of the medical dataset shown in Table 5.1 with respect to both the demographics and the diagnosis codes of the patients.

In what follows, we discuss some popular, non-commercial systems for performing data anonymization and highlight their functionality and limitations, when applied to medical data. Subsequently, we provide a discussion of concepts that are necessary to understand how the anonymization of relational, transactional, and RT-datasets is performed. Last, we provide an overview of SECRETA, a system that enables healthcare data owners to effectively anonymize their medical datasets. The system allows to evaluate and compare a range of anonymization methods, with respect to their achieved level of privacy, utility, and computational runtime.

5.2 Related Work

Data anonymization tools are important both for allowing users to experiment with different anonymization methods and understand their workings, as well as for enabling non-experts to anonymize their data. However, only a limited number of non-commercial data anonymization tools are currently available. Below, we discuss three of the most feature-rich non-commercial tools for data anonymization.

The Cornell Anonymization Toolkit (CATtool) [23] supports k -anonymity and ℓ -diversity [12] on relational datasets, essentially offering protection from identity and sensitive attributes' disclosure. The CATtool enables users to select an input dataset, specify the attributes which act as quasi-identifiers, select any sensitive attributes, and finally set the k or ℓ parameter to a desired value. The tool then applies the selected anonymity method and presents in two graphs the distribution of the original and anonymized values for the quasi-identifying and the sensitive attributes. Moreover, the tool calculates a risk value for the dataset, based on the probability of an attacker to accurately link a user to a sensitive attribute value.

Next, the TIAMAT tool [2] supports the k -anonymity model [20] on datasets containing only relational attributes, while the authors have stated that they plan to extend it to also support ℓ -diversity [12] and t -closeness [10] in a future release. TIAMAT has the following interesting features:

- It allows users to visually create and edit the taxonomies of relational attributes, and also select which attributes in the dataset will act as quasi identifiers.
- It enables users to evaluate the performance of the supported methods, measuring the *NCP* [24] and the efficiency, and presenting them in graphs.
- It supports the visual representation of several executions of the algorithm with variable values of parameter k . This mode is called *head-to-head analysis*, and allows users to obtain a better understanding of the effect of parameter k in information loss and efficiency.

Prasser et al. recently introduced ARX [19], an interesting tool for anonymizing relational datasets. This tool supports algorithms enforcing k -anonymity [20], ℓ -diversity [12], t -closeness [10] and δ -presence [13]. ARX has a user-friendly editor for generalization hierarchies, where users can construct hierarchies to meet their requirements. It also supports a number of utility measures, such as monotonic and non-monotonic variants of height, precision, discernability and entropy. It presents the anonymized dataset to the end-user and reports the above metrics in a graphical representation. Another interesting feature of ARX is that it can be easily extended; the authors provide an API which enables users to implement and integrate to the tool their own methods, or use the tool with their existing software solutions.

The SECRETA system, which will be presented in detail in the remainder of this chapter, supports the *core* functionalities of CATtool, TIAMAT and ARX, as can be seen in Table 5.3, although it currently lacks an anonymization API as well as the advanced data visualization and re-identification risk analysis capabilities offered

Table 5.3 Comparison of data anonymization tools

	CATtool	TIAMAT	ARX	SECRETA
Rel. datasets	✓	✓	✓	✓
Trans. datasets				✓
RT-datasets				✓
Privacy constraints				✓
Utility constraints				✓
Automatic hierarchies			✓	✓
Distributions plot			✓	✓
NCP/UL/ARE plots		✓		✓
Anonymization API			✓	
Privacy risk analysis			✓	
Advanced visualization			✓	

by ARX. In more detail, SECRETA supports the Incognito [8], the Mondrian [9], the Cluster [14], the Top-down [4], and the Full subtree bottom-up anonymization algorithms for relational datasets. For transaction/set-valued datasets, SECRETA supports the COAT [11], the PCTA [5], the Apriori, the LRA [21], and the VPA [21] algorithms. Furthermore, SECRETA enables the combined use of relational and transaction anonymization algorithms to anonymize RT-datasets. This is achieved with the help of three methods, called **RMERGE**, **TMERGE**, and **RTMERGE** [14], which combine a relational with a transaction anonymization algorithm.

5.3 Overview of SECRETA

This section describes the components of SECRETA, which we broadly divide into *frontend* and *backend* components. The frontend offers a Graphical User Interface (GUI), which enables users to issue anonymization requests, visualize and store experimental results. The backend consists of components for servicing anonymization requests and for conducting experimental evaluations. The architecture of SECRETA is illustrated in Fig. 5.1.

5.3.1 Frontend of SECRETA

The frontend of SECRETA is implemented using the QT framework (available at: <https://qt-project.org>). Using the provided GUI, users are able to:

1. Select relational, transaction or RT-datasets for anonymization.
2. Specify generalization hierarchies and query workloads.
3. Select and configure data anonymization algorithms.

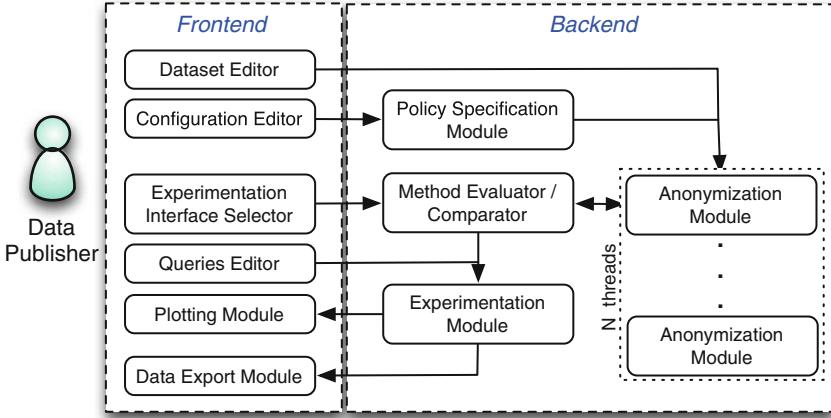


Fig. 5.1 System architecture of SECRETA

4. Execute experiments and visualize the experimental results.
5. Export anonymized datasets and experimental results, in a variety of formats.

In what follows, we elaborate on the components of the frontend.

Dataset Editor The dataset editor enables users to select datasets for anonymization. The datasets can have relational and/or transaction attributes, and they need to be provided in a Comma-Separated Values (CSV) format. Relational datasets consist of records (rows) and attributes (columns), where each record has a specific value for each attribute. Attributes (respectively, values) may be numerical or categorical. Transaction datasets consist of records (rows), where each record is associated with a set of items (literals). Different records may contain a different number of items, which are provided as a comma-separated list. RT-datasets consist of a set of relational attributes and one transaction attribute.¹ Once a dataset is loaded to the dataset editor, the user can modify it (edit attribute names and values, add/delete rows and attributes, etc.) and store the changes. The user can also generate data visualizations, such as histograms of attributes. Figure 5.2 shows a loaded dataset and some of the supported data visualizations.

Configuration Editor The configuration editor allows users to select hierarchies and to specify privacy and utility policies. Hierarchies are used by all anonymization algorithms, except from COAT [11] and PCTA [5], whereas privacy and utility policies are only used by these two algorithms to model such requirements.

¹We note that in the case of RT-datasets containing more than one transaction attributes, these attributes can be integrated into a single transaction attribute by changing their domain and values' specification. Assuming two such attributes X and Y , with X having values x_1, x_2 and Y values y_1, y_2 , we can replace these attributes with a single transaction attribute, say Z , with domain $\{X.x_1, X.x_2, Y.y_1, Y.y_2\}$.

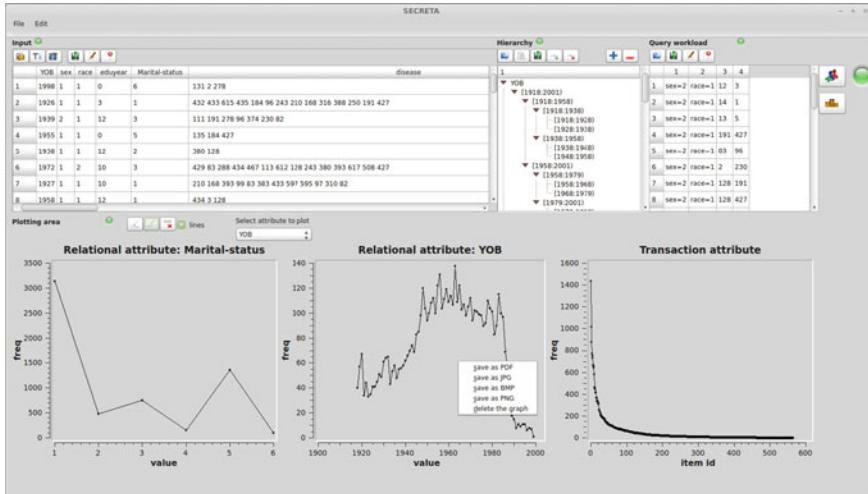


Fig. 5.2 The main screen of SECRETA

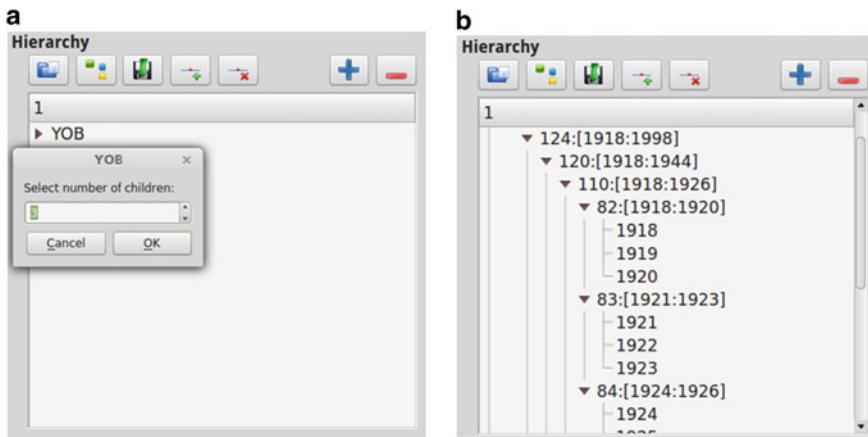


Fig. 5.3 Automatic creation of hierarchies. (a) Selecting the number of splits per level of the hierarchy and (b) Displaying the produced hierarchy

Hierarchies and policies can be uploaded from a file, or automatically derived from the data, using the implemented algorithms from [21]. Figure 5.3 shows an example of the automatic hierarchies creation process. For instance, in Fig. 5.3a the user selects 3 as the number of splits per hierarchy-level for relational attribute age. Thus, for attribute age, every parent node in the hierarchy will have three children nodes (e.g., parent [1918 : 1926] has children [1918 : 1920], [1921 : 1923] and [1924 : 1926]). The created hierarchy is depicted in Fig. 5.3b.

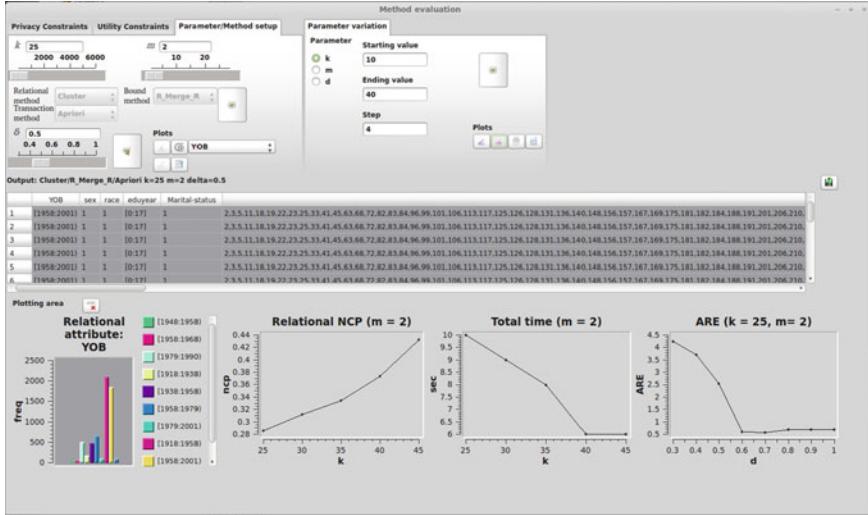
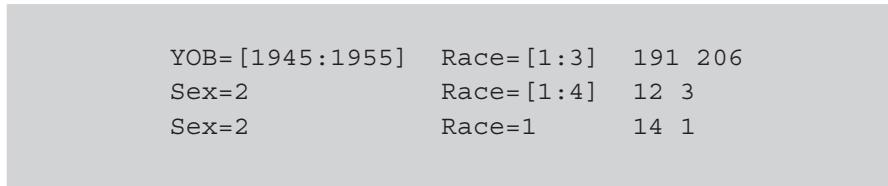


Fig. 5.4 The evaluation mode: method evaluation screen of SECRETA

Queries Editor This component allows specifying query workloads, which will be used to evaluate the utility of anonymized data in query answering. SECRETA supports the same type of queries as [24], and uses Average Relative Error (ARE) [24] as a defacto utility indicator. The query workloads can be loaded from a file and edited by the user, or can be loaded directly using the GUI.



In the above queries workload, for instance, the first query selects all users in the dataset with YOB (year of birth) in [1945:1955], Race in [1:3], and having the diagnosis codes 192 and 206.

Experimentation Interface Selector This component sets the operation mode of SECRETA. Figure 5.4 presents the Evaluation mode, in which users can evaluate a selected anonymization algorithm, while Fig. 5.5 shows the Comparison mode, which allows users to compare multiple anonymization algorithms. Through these interfaces, users can select and configure the algorithm(s) to obtain the anonymized data, store the anonymized dataset(s), and generate visualizations that present the performance of the algorithm(s).



Fig. 5.5 The comparison mode: methods comparison screen of SECRETA

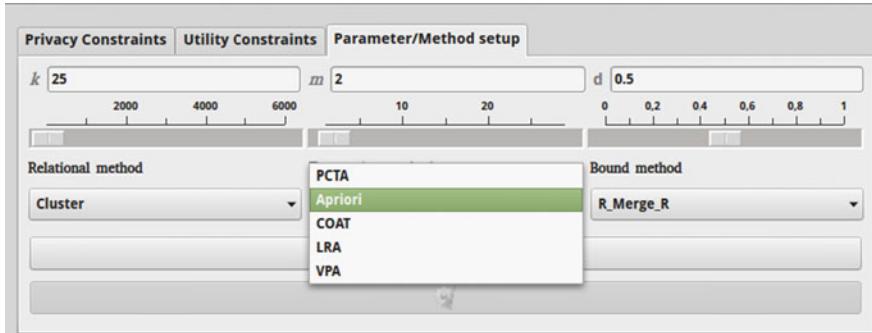


Fig. 5.6 The experimentation interface selector

Figure 5.6 highlights the algorithms' configuration pane. In this figure, the user has set parameters $k = 25$, $m = 2$ and $\delta = 0.5$, and has selected the Cluster algorithm as the relational anonymization method, the Apriori algorithm as the transactions anonymization method, and RMERGE_r as the bounding method, in order to anonymize an input RT-dataset.

Plotting Module The plotting module is based on the QWT library (<http://qwt.sourceforge.net/>) and supports a series of data visualizations that help users analyze their data and understand the performance of anonymization algorithms, when they are applied with different configuration settings. Specifically, users can visualize information about:

1. The original and the produced anonymized datasets (e.g., plot histograms of selected attributes, plot the relative differences in the frequencies between original and generalized values, etc.)
2. The anonymization results, for *single* and *varying* parameter execution.

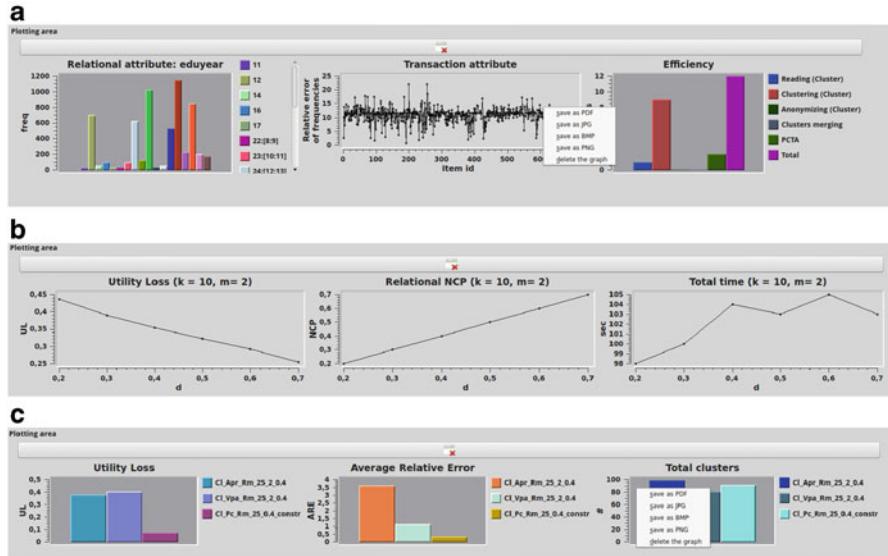


Fig. 5.7 Plots for (a) the original dataset, (b) varying parameters execution, and (c) the comparison mode

In the single parameter execution, the results are derived with fixed, user-specified parameters and include frequencies of generalized values in relational attributes or transaction attributes, runtime, etc. In varying parameter execution, the user selects the start/end values and step of a parameter that varies, as well as fixed values for the remaining parameters. The plotted results include data utility indicators and runtime vs. the varying parameter.

Data Export Module The data export module allows exporting datasets, hierarchies, policies, and query workloads, in CSV format, as well as graphs in PDF, JPG, BMP or PNG format.

Various instances of the plotting and data export modules are shown in Fig. 5.7. Specifically, Fig. 5.7a presents a histogram of the generalized values of attribute `eduyear`, the relative error between the frequencies of the items in the transaction attribute (comparing the frequency of each item in the original vs. the anonymized dataset), and the runtime of the various phases of the selected algorithm (efficiency). Figure 5.7b presents the UL in the transaction part of the data, for fixed values of k and m and varying parameter δ , as well as the NCP and the runtime of Algorithm **RMERGE**, for different values of parameter δ . Figure 5.7c shows the UL, the ARE and the total number of created clusters for the three different executions, where:

1. Cluster was used for the relational part, Apriori for the diagnosis codes, **RMERGE**, as the bounding method, and $k = 25$, $m = 2$ and $\delta = 0.4$.
2. Cluster was used for the relational part, VPA for the diagnosis codes, **RMERGE**, as the bounding method, and $k = 25$, $m = 2$ and $\delta = 0.4$.

3. Cluster was used for the relational part, Apriori for the diagnosis codes, RMERGE_r as the bounding method, $k = 25$, $\delta = 0.4$ and the specified utility and privacy constraints.

These methods and their parameterisation are presented in Sect. 5.3.3. Figure 5.7a–c help users to conclude that the third configuration is superior to the others.

5.3.2 Backend of SECRETA

In this section we present the backend of SECRETA. First, we provide some key definitions that are necessary to introduce the methods and utility metrics implemented into SECRETA (for more information please refer to [6, 7]). Then, we discuss the specific modules that have been implemented in the backend.

5.3.2.1 Key Definitions

In Sect. 5.1 we highlighted the importance of medical data anonymization algorithms. This need comes from the ability of attackers to use their knowledge about specific patients’ quasi-identifying values in order to uniquely identify the corresponding patients’ records in the dataset. As attackers can hold partial knowledge about various patients, it is critical to protect the data from all possible attackers.

For instance, an attacker can identify a patient in a relational dataset, using some known values over the patient’s quasi-identifiers. In a transactions dataset, it is reasonable to consider that an attacker may know up to m diagnosis codes of the patient, where m can be arbitrarily large, even covering the entire domain. Finally, in a RT-dataset an attacker may know the patient’s set of quasi-identifiers for the relational part and up to m of the patients’ diagnosis codes (transaction part). We note that this type of adversarial knowledge is more powerful than the one considered in the cases of relational or transaction datasets.

Relational Datasets In relational datasets, the most commonly used anonymity model is k -anonymity [20]. k -anonymity protects from identity disclosure, as it requires that a record becomes indistinguishable from at least $k - 1$ other records in the dataset. This group of at least k similar records is called an *equivalence class*.

Most of the methods satisfying k -anonymity by employing data *generalization* or *suppression*. These operations are applied to the quasi-identifier values. Specifically, data generalization replaces the original (crisp) values with more general values, and suppression deletes the corresponding values.

The values of a relational attribute can be either *categorical* (like gender) or *numerical* (like age). Generalizing a categorical value, simply replaces this value with a more general but semantically close one, using a generalization hierarchy. For example, in the first record of Table 5.2a, the original country’s value France

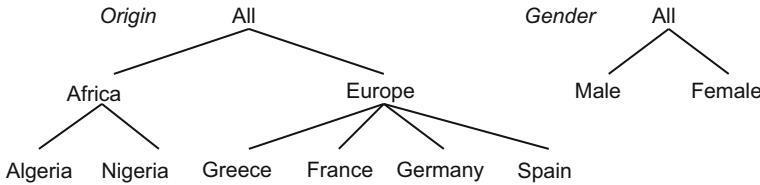


Fig. 5.8 An example of a hierarchy tree

is replaced with the more general value *Europe* according to the hierarchies of Fig. 5.8. Similarly, a numerical value can be replaced with an interval of consecutive values, where the original value belongs to the interval. For instance, the original value of *age* on record 1 in Table 5.2a can be replaced with [19:22].

These are the possible schemes for methods employing generalization:

Full-domain generalization [8, 20]. In this scheme, the same level of a hierarchy tree is used for all the values of an attribute. For example, using the hierarchies of Fig. 5.8, if *Origin*'s value *Algeria* was generalized to *Africa*, then it would be also required that *Nigeria* is generalized to *Africa* and *Greece*, *France*, *Germany* and *Spain* are generalized to *Europe*. This scheme has the smallest search space for a possible solution, as the anonymization algorithm climbs rapidly the hierarchy tree level-by-level, entailing, on the other hand, significant data distortion.

Subtree generalization [4, 14]. In this scheme, at a node, either all values are generalized to their parent, or none is generalized. In more detail, assuming the hierarchy of Fig. 5.8, if *Algeria* is generalized to *Africa*, then also *Nigeria* has to be generalized to *Africa*, but *Greece*, *France*, *Germany* and *Spain* can maintain their values. This scheme has an increased search space for finding a solution, but typically results in a lower level of data distortion when compared to the previous scheme.

Multidimensional generalization [9]. In the previous schemes, when a value is generalized, all its instances are generalized. Such schemes are called *global recoding* schemes. Alternatively, the multidimensional generalization or *local recoding scheme*, allows the generalization of some instances of a value, while other instances may remain to their original values or be generalized in different levels of the hierarchy. In more detail, this scheme enables two groups of records, having the same value on a quasi-identifier, to be generalized independently into different levels of the hierarchy. For example, let us assume two groups of records. The first group has two records, both sharing the same original value *Nigeria*. The second group has also two records, having values *Nigeria* and *Algeria*, respectively. The first group is already 2-anonymous with respect to attribute *Origin*, and thus no generalization has to be applied. On the other hand, in the second group, we need to generalize both values to *Africa* to make it 2-anonymous. This scheme enables us to generalize only the groups of records that violate the anonymity principle, producing less distorted data than the full-domain and subtree generalization schemes.

Table 5.4 A $(2, 2^2)$ -anonymous dataset with privacy constraints $\mathcal{P} = \{\text{Flu}, \text{Herpes}\}$ and utility constraints $\mathcal{U} = \{\{\text{Asthma}, \text{Flu}\}, \{\text{Herpes}, \text{Eczema}\}\}$

Id	Age	Origin	Gender	Disease
1	[19:22]	Europe	Male	Asthma Flu (Herpes, Eczema)
2	[19:22]	Europe	Male	Asthma Flu (Herpes, Eczema)
3	[28:39]	Europe	Female	(Asthma, Flu) (Herpes, Eczema)
4	[28:39]	Europe	Female	(Asthma, Flu) (Herpes, Eczema)
5	[55:70]	Africa	All	Asthma
6	[55:70]	Africa	All	Flu

Transaction Datasets For the case of transactions datasets, the most commonly employed model is k^m -anonymity [5, 11, 21], which considers attackers that know up to m diseases (generally *items*) associated with a patient, and guarantees that there are least k records in the dataset sharing the same set of m diseases.

The methods that achieve k^m -anonymity can be categorized as follows:

Hierarchy based generalization [21]. This method applies generalization using a predefined hierarchy, which groups together semantically close items, e.g., diagnosis codes. For example, Essential hypertension (ICD-9 code: 401) could be generalized to Hypertensive disease (ICD-9 codes: 401-405) and then to Diseases of the circulatory system (ICD-9 codes: 390-459), if necessary.

Constraint based generalization [5, 11]. This approach does not use a hierarchy, but instead groups together items, e.g., diagnosis codes, based on their semantics. The corresponding methods can select only certain items to act as quasi-identifiers or enable data publishers to force certain items to become generalized together. Thus, they entail less distortion than hierarchy based methods, and also produce results that are more meaningful in certain medical applications. In Table 5.4 we present a $(2, 2^2)$ -anonymous version of the dataset of Table 5.1. In this case we consider as potentially linkable the combination of Flu and Herpes. We also enable our method to generalize together only Asthma with Flu, and Herpes with Eczema.

Typical anonymity models consider all diagnosis codes as potentially linkable, which may not be the case in the medical domain. For that reason, SECRETA allows data owners to specify an arbitrarily large set of *privacy constraints*, each holding a combination of items that are considered as risky and potentially linkable. The anonymization algorithm takes as input the user-specified privacy constraints and applies the necessary data transformations to guarantee that each constraint is satisfied in the resulting dataset. In this way, each individual (record) is linked to a sufficiently large number of records with respect to the set of items (a.k.a. itemsets) specified in the privacy constraints.

In more detail, a *privacy constraint* is a non-empty set of items that are specified as potentially linkable. The union of all privacy constraints formulates a *privacy constraints set* or a *privacy policy*. In order to satisfy a privacy policy, each

privacy constraint in the policy has to be independently satisfied. For a parameter k , a privacy constraint is satisfied if:

1. The corresponding itemset (of the privacy constraint) appears in at least k transactions of the anonymized dataset, *or*
2. It does not appear in the original dataset and each of its proper subsets either appears in at least k transactions in the anonymized dataset, or it does not appear in the original dataset.

For example, in Table 5.4, records 5 and 6 are considered safe with respect to the given privacy policy \mathcal{P} , as only records holding together Flu and Herpes are considered unsafe. We note, however, that the user could have also specified Flu and Herpes as individual privacy constraints.

Another important factor in medical datasets, is the ability to select which data generalizations are acceptable, in order to produce data that is meaningful for supporting medical analyses. In order to enforce only such *eligible* data generalization actions, a set of *utility constraints* (collectively known as a *utility policy*) is supported by selected anonymization algorithms. More formally, a utility constraint declares the set of items that are allowed to be generalized together. The set of all utility constraints defines the *utility policy*. A utility policy is satisfied if:

1. Any generalization involves only items (e.g., diagnosis codes) that belong to the same utility constraint, *and*
2. The fraction of suppressed items (e.g., diagnosis codes) in the anonymized dataset is at most $s\%$ (where s is a value specified by the data owner to reduce the number of item suppressions in the dataset).

An example is presented in Table 5.4, where Herpes is generalized together with Eczema, and Asthma with Flu, thereby satisfying the utility policy \mathcal{U}_c .

RT Datasets To anonymize RT-datasets the privacy model of (k, k^m) -anonymity has been defined [14]. This privacy model aims to protect users from attackers whose knowledge spans between the relational attributes and the items of the users. (k, k^m) -anonymity guarantees that there exist groups of records with size at least k , where each group holds patients with the same quasi-identifiers' values and the same set of m items (e.g., diagnosis codes). (k, k^m) -anonymity can be achieved through the framework described in [14]. In this framework, a data publisher initially creates clusters of records having size between $[k, 2k - 1]$. These clusters can be created to support either records (a) with semantically close relational attributes, or (b) semantically close items. Then, a data publisher can use the methods described in Sect. 5.3.2.1 to anonymize the relational part of the data. Finally, inside each cluster, the methods of Sect. 5.3.2.1 are applied to guarantee privacy for the items (e.g., diagnosis codes) of the clusters.

Utility Metrics As there are numerous methods that can be applied to anonymize a dataset, data publishers need access to intuitive and accurate metrics, in order to evaluate the information loss of each method. In relational datasets, the information loss of a generalization may be captured by using the *Normalized Certainty Penalty*

(NCP)[24] metric. NCP is calculated by taking under consideration the hierarchy for categorical values and the generalized space for numerical attributes.

In more detail, suppose that a record t has value u on a categorical attribute A_i , where u is replaced with the set of values $\tilde{v} = \{u_1, \dots, u, \dots, u_l\}$ during generalization. The *Normalized Certainty Penalty* for categorical attribute A_i in record t is defined as:

$$NCP_{A_i}(t) = \begin{cases} 0, & |\tilde{v}| = 1 \\ |\tilde{v}|/|R|, & \text{otherwise} \end{cases}$$

where $|\tilde{v}|$ is the number of distinct values in set $\{u_1, \dots, u, \dots, u_l\}$ and $|A_i|$ is the number of distinct values of the attribute A_i .

The NCP for a numerical attribute A_i , which is generalized to interval $[x:y]$ in record t is defined as:

$$NCP_{A_i}(t) = \frac{|y - x|}{|max_{A_i} - min_{A_i}|},$$

where max_{A_i} (min_{A_i}) is the maximum (minimum) value of attribute A_i in dataset D . The NCP of a record t is calculated as:

$$NCP(t) = \frac{\sum_{i=1}^n NCP_{A_i}(t)}{n},$$

where $NCP_{A_i}(t)$ is the NCP of attribute A_i in record t , and n is the total number of attributes in a record.

Consequently, the total NCP of an anonymous dataset D is defined as:

$$NCP(D) = \frac{\sum_{t \in D} \sum_{i=1}^n NCP_{A_i}(t)}{|D|},$$

where $|D|$ is the number of records of dataset D .

In order to capture the information loss in transaction datasets, we use *Utility Loss (UL)* [5, 11]. UL calculates information loss, by assigning a weight to each generalized item, which reflects the semantics of the diagnosis codes that are generalized together. UL was first proposed by Loukides et al. [11], for the case of global recoding. Then, it was adapted for local recoding by Pouli et al. [14]. The latter definition is more suitable for the anonymization of RT-datasets, because it calculates the information loss per cluster. For a generalized item \tilde{u} , a record t , and a dataset D , *UL* is defined as follows:

$$UL(\tilde{u}) = (2^{|\tilde{u}|} - 1) \cdot w(\tilde{u}),$$

$$UL(t) = \frac{\sum_{\forall \tilde{u} \in t} UL(\tilde{u})}{2^{\sigma(t)} - 1},$$

$$UL(D) = \frac{\sum_{\forall r \in D} UL(r)}{|D|},$$

where $|\tilde{u}|$ is the number of items mapped to \tilde{u} , $w(\tilde{u}) \in [0, 1]$ is a weight reflecting the importance of items \tilde{u} respectively [11], and $\sigma(t)$ is the sum of sizes of all generalized items in record t .

Last, *Average Relative Error* (ARE) [9] can be used to quantify data utility in both the relational and the transaction part of a record. ARE calculates the average number of records that are retrieved incorrectly, as part of the answer set to a workload of COUNT queries posed on the anonymized dataset.

As an example, consider the following query on relational data:

```
QUERY1= SELECT COUNT(*)
          FROM dataset
          WHERE Age=19 AND Origin=France.
```

When the above query is issued on the dataset of Table 5.1 the result is 1, as only one patient has Age=19 and Origin=France. However, when the same query is issued to the dataset of Table 5.2a, there are two possible answers (i.e., patients 1 and 2 having Age=[10:22] and Origin=Europe).

Similarly, for the following query issued on transaction data:

```
QUERY2= SELECT COUNT(*)
          FROM dataset
          WHERE Disease=Asthma AND Disease=Herpes.
```

When this query is issued on the dataset of Table 5.1 the result is 1, as only the patient of the first record has diseases Asthma and Herpes. When the same query is issued to the dataset of Table 5.2b, there are three possible answers (i.e., patients 1, 2 and 4 have the generalized item {Herpes, Eczema}). Based on the semantics of the applied generalization, a patient having {Herpes, Eczema} may suffer either from Herpes, or Eczema, or both.

5.3.3 Components

The backend of SECRETA is implemented in C++ and consists of four components, namely the *Policy Specification module*, the *Method Evaluator/Comparator*, the *Anonymization module*, and the *Experimentation module*. In a nutshell, SECRETA invokes one or more instances of the anonymization module with the specified algorithm and parameters. The anonymization results are collected by the Method Evaluator/Comparator component and forwarded to the Experimentation module.

From there, results are forwarded to the Plotting module for visualization, and/or to the Data Export module for data export.

Policy Specification Module The Policy Specification module invokes algorithms that automatically produce hierarchies [22], as well as the strategies in [11], which generate privacy and utility policies. The hierarchies and/or policies are then used by the anonymization module.

Method Evaluator/Comparator Module This module implements the functionality that is necessary for supporting the interfaces of the Evaluation and the Comparison modes. Based on the selected interface, anonymization algorithm(s) and parameters, this component invokes one or more instances (threads) of the anonymization module. After all instances finish, the evaluator/comparator component collects the anonymization results and forwards them to the Experimentation module.

Anonymization Module The Anonymization module is responsible for executing an anonymization algorithm with the specified configuration. SECRETA supports nine algorithms in total. In more detail, the following five algorithms are applicable to datasets with relational attributes.

1. **Incognito** [8] is an algorithm that enforces k -anonymity on relational data by applying full-domain generalization. Incognito works in an apriori fashion, initially generalizing single attributes and then larger combinations of attributes in the set of quasi-identifiers, until it produces an anonymous dataset. Incognito minimizes the information loss in anonymized data by selecting generalizations that produce more equivalence classes.
2. **Cluster** [14] is an algorithm that employs a multidimensional generalization scheme achieving minimal information loss, entailing only a minimal efficiency overhead. Initially, Cluster selects a random record as a seed and finds $k - 1$ other records with similar relational attributes' values to that seed, formulating a cluster of size k . This intuitive selection of the cluster's records, typically results in very low NCP scores.
3. **Mondrian** [9] is a popular method that uses a multidimensional generalization scheme. It is a top-down specialization algorithm, which selects a quasi-identifying attribute and performs local recoding specialization.
4. **Top-down** [4] is another specialization algorithm. It performs specialization to one quasi-identifying attribute at a time, as Mondrian does, but instead it uses a global recoding scheme.
5. **Full subtree bottom-up** [14] is an algorithm that traverses the hierarchy tree from bottom to top, generalizing every attribute in its parental value. This method outperforms full-domain generalization methods, as it uses an extended search space to find a solution.

As can be seen, these algorithms differ with respect to the way they transform quasi-identifier values and also the way they operate. Mondrian is the fastest of these algorithms and Full subtree bottom-up generalization typically achieves better

utility. However, each of these algorithms has features that may be important in particular applications. For example, Incognito can easily generate all possible generalizations that can be generated according to its recoding model.

SECRETA supports the following transaction anonymization algorithms:

1. **COAT** [11] is an algorithm that prevents identity disclosure. This algorithm employs the constraint-based generalization model and can support k^m -anonymity, as well as privacy-constrained anonymity. It also supports utility constraints.
2. **PCTA** [5] works in a similar manner to COAT, supporting both k^m -anonymity and privacy-constrained anonymity, as well as utility constraints. PCTA uses clustering-based heuristics and is significantly more effective than COAT, reducing utility loss. However, PCTA is less efficient than COAT.
3. **Apriori Anonymization** [22] is a popular method that has provided the basis for many other anonymization algorithms [15, 16, 18, 22]. This method enforces k^m -anonymity, initially fixing single problematic diagnosis codes and then moving up to sets of problematic diagnosis codes with sizes $2, \dots, m$. It uses hierarchy-based generalization based on a user-provided generalization hierarchy.
4. **LRA** [22] and **VPA** [22] are heuristics based on Apriori Anonymization, employing hierarchy based local recoding generalization. These algorithms initially partition the dataset in clusters with similar items and then generalize the items inside each cluster to enforce the k^m -anonymity principle.

All these algorithms are heuristics, which differ according to the way they generalize data and their objective functions. It is important to support these algorithms, because none of them is superior to others in all applications. For example, COAT is more effective than Apriori Anonymization, but it is not as scalable.

Additionally, SECRETA supports three bounding methods [14], which merge together clusters, until the total NCP of the dataset reaches a user defined threshold δ . These methods can be applied to datasets that have been anonymized using Incognito, Cluster, TDS and Full subtree bottom-up algorithms. Note that Mondrian is not supported, as it does not use a hierarchy to construct equivalence classes. The supported bounding methods are:

1. **RMERGE_r**. This method merges together clusters based on their respective relational attributes. It uses the NCP metric to evaluate how close two clusters are, and chooses the clusters that produce the lowest NCP.
2. **TMERGE_r**. This method merges together clusters based on their respective transaction items. Specifically, it evaluates how many common, different and total diagnosis codes exist between every pair of records between the two clusters, and chooses the clusters that have the most semantically closest items.
3. **RTMERGE_r**. This method takes into consideration both the relational and the transaction parts of the records inside each cluster. Its performance lies between **RMERGE_r** and **TMERGE_r**.

An overview of the supported algorithms in SECRETA is presented in Table 5.5.

Table 5.5 Overview of the methods supported from SECRETA

	Relational datasets	Transaction datasets	RT-datasets
Incognito	✓		✓
Mondrian	✓		
Cluster	✓		✓
Top-down	✓		✓
Bottom-up	✓		✓
PCTA		✓	✓
COAT		✓	✓
Apriori		✓	✓
LRA		✓	✓
VPA		✓	✓

Experimentation Module This module is responsible for producing visualizations of the anonymization results and of the performance of the anonymization algorithm(s), in the case of *single* and *varying* parameter execution. For visualizations involving the computation of ARE, input is used from the Queries Editor module. The produced visualizations are presented to the user, through the Plotting module, and can be stored to disk using the Data Export module.

5.4 Using SECRETA

We will now demonstrate all the necessary steps that a user has to complete in order to anonymize and evaluate a dataset. Let us consider a user (e.g., a healthcare facility) holding a medical dataset, which contains the demographics and a set of diagnosis codes for various patients. Let us also assume that the user wants to anonymize and publish this dataset, enabling medical researchers to study the correlation between the age and sex of a patient with certain diagnosis codes.

To validate that the anonymized dataset can still be used for meaningful analysis, the user wants to issue some COUNT queries on the original and the anonymous datasets, and measure the average relative error.

- Prepare the dataset in order to be compatible with SECRETA.
- Load the dataset in SECRETA.
- Edit the dataset using the embedded dataset editor of SECRETA.
- Define a hierarchy for the demographic attributes.
- Load and edit a workload of COUNT queries.
- Perform a set of data visualizations to the original dataset.
- Compare the available anonymization methods for the dataset in order to select the most appropriate one.
- Anonymize the dataset using the selected method.
- Export various graphs reporting on a set of quality and efficiency metrics.

Table 5.6 (a) A fictitious medical dataset presenting the Year-Of-Birth, Sex, Race, and Diagnosis codes of patients, (b) mapping of attribute Sex, and (c) mapping of attribute Race

YOB	Sex	Race	Diagnosis codes
1998	1	1	131 2 278
1926	1	1	432 433 615 435 184 96 243 210 168 316 388
1955	1	1	135 184 427
1938	1	1	380 128
1972	1	2	429 83 288 434 467 113 612 128 243 380 393

(a)

id	Sex
1	Male
2	Female

(b)

id	Race
1	White
2	Black
3	American Indian/Alaska Native
4	Asian
5	Native Hawaiian/Pacific Islander
6	Multiple races reported

(c)

5.4.1 Preparing the Dataset

We consider that the user has a medical dataset with information like the one presented in Table 5.6a. In Table 5.6b, c, we present the mapping for attributes `sex` and `race` respectively. For instance, `sex=1` in the original dataset, means that this patient is male. In order to make this dataset compatible with SECRETA, the user has to export it to a file of the following format:

```
//YOB,sex,race,diagnosis_codes
1998,1,1,0,6, 131 2 278
1926,1,1,3,1, 432 433 615 435 184 96 243 210 168 316 388
1955,1,1,0,5, 135 184 427
1938,1,1,12,2, 380 128
1972,1,2,10,3, 429 83 288 434 467 113 612 128 243 380 393
```

The first line holds the name of all attributes of the dataset and is optional. In the following lines, the relational attributes have to be separated by comma, while the items of the transaction attribute have to be separated by spaces. Finally, a comma and a space has to be used in order to separate the relational attribute from the transaction attribute.

RT-dataset					
	YOB	sex	race	diagnosis_codes	
5	1938	1	1	380	128
6	1972	1	2	429	83 288 434 467 113 612 128 243 380 393 617 508 427
7	1927	1	1	210	168 393 99 83 383 433 597 595 97 310 82
8	1958	1	1	434	3 128
9	1960	1	1	230	
10	1932	1	1	431	485 427
11	1941	1	4	393	83 388 191 96
12	1943	1	1	458	599 191
13	1954	2	1	393	273 392 156 323 41 80 148 128 427 230
14	1934	2	1	80	

Fig. 5.9 The dataset editor

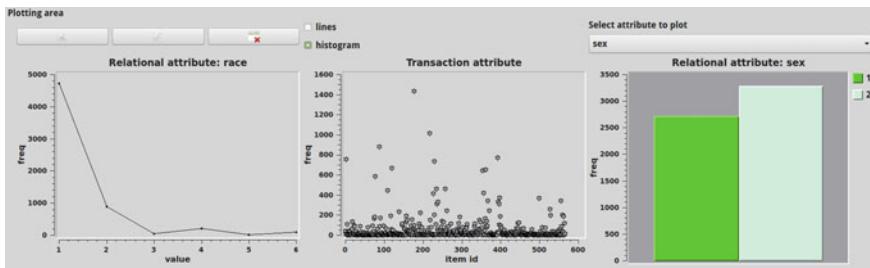


Fig. 5.10 Frequency plots of the original dataset

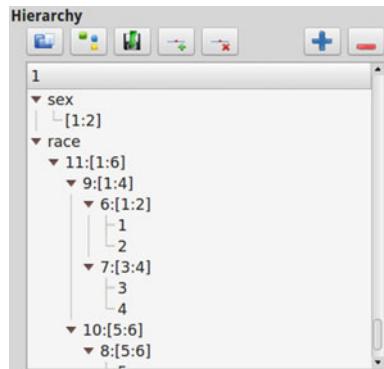
5.4.2 Using the Dataset Editor

The user can now load the dataset in the *dataset editor* using the button. SECRETA automatically recognizes the type and the format of the dataset, and presents the loaded dataset (Fig. 5.9).

After that, the user is able to edit the attribute names of the dataset, by double clicking on their respective title column. The user can also edit the values of a record, select and delete several records, or insert new records.

Finally, the user may overwrite the existing dataset with the modified one, or export it to a file. Subsequently, the user can analyze the dataset by plotting histograms of the frequency of values in any attribute on the bottom pane of the main screen of SECRETA (see Fig. 5.10).

Fig. 5.11 The hierarchy specification area



5.4.3 The Hierarchy Editor

Next, the user can define a hierarchy for the relational attributes of the dataset. This step is applicable only on relational and RT-datasets. A hierarchy can be either loaded from a file or be automatically generated by SECRETA, using the automatic hierarchy generation feature. This feature is triggered by clicking the button. When this feature is used, the user inputs the number of children that a leaf node shall contain, for every attribute of the dataset. The generated hierarchy for the example dataset is presented in Fig. 5.11. This hierarchy is fully browsable and editable, and can be exported to a text file.

5.4.4 The Queries Workload Editor

The user can also load a query workload from a file, in order to issue the desired COUNT queries in the original and anonymized datasets. Each COUNT query has to be in the following format:

```
attribute=<range of values> ... diagnosis_code_id ...
```

In more detail, the user enters the name of a relational attribute and a desired value or range of values. The user can enter more than one relational attributes in the same query. Next, the user enters the desired items for the transaction attribute. For example, a valid query workload is:

```
YOB=[1978:1985] sex=1 12 92
YOB=1978 sex=[1:2] 43 92
YOB=1978 sex=1 12 43
```

In this example, the first query counts all male patients having YOB (year-of-birth) between 1978 and 1985, and being diagnosed with diseases 12 and 92.

After loading the query workload (by pressing the  button on the upper right corner of SECRETA), the user can edit single queries, add or remove queries, and finally save the query workloads to a file.

5.4.5 Evaluating the Desired Method

Having prepared the dataset, defined a hierarchy and a queries workload, the user can now enter the Method Evaluation interface, by clicking the  button. This interface enables the user to configure, apply, and evaluate a single method.

Initially, the user can use the Method Evaluation interface and set the values for parameters k , m , δ , using the text boxes or the corresponding sliders (Fig. 5.12). Notice that the bounds for parameters k and m are automatically set by the system based on the input dataset. For instance, if the users' dataset contains 6000 records and no record has more than 29 values, k has allowable values in $[2, 6000]$ and $m \in [1, 29]$. By definition, $\delta \in [0, 1]$ and the user can specify its exact value. Then, the user may select three algorithms, one for anonymizing the relational attributes, one for the transaction attribute, and one as the bounding method.

Next, the user may initiate the anonymization process. When this process ends, a message box with the summary of results will be presented (Fig. 5.13) and the anonymized dataset will be displayed in the output area (Fig. 5.14). Last, the user can select a number of data visualizations. These visualizations will be presented in the plotting area (Fig. 5.15) and may illustrate any combination of the following:

1. ARE scores for various parameters (e.g., for varying δ and fixed k and m , or for varying k and fixed δ and m , etc.)
2. The time needed to execute the overall algorithm and its different phases.
3. The frequency of all generalized values in a selected relational attribute.
4. The relative error between the frequency of the transaction attribute values (items), in the original and in the anonymized dataset.

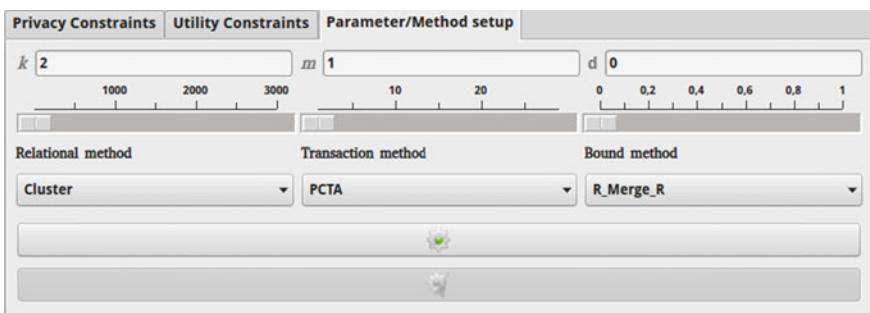


Fig. 5.12 Method parameters setup

Fig. 5.13 A messagebox with the results summary

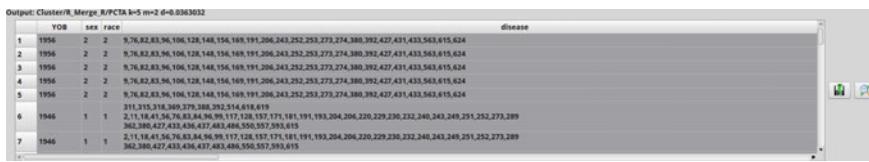
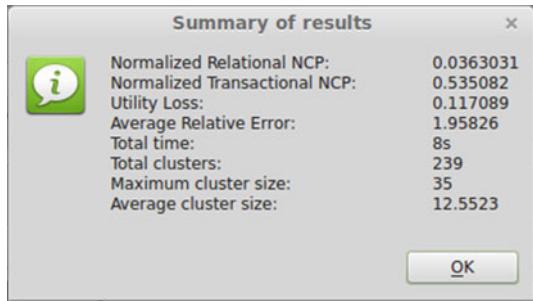


Fig. 5.14 The data output area

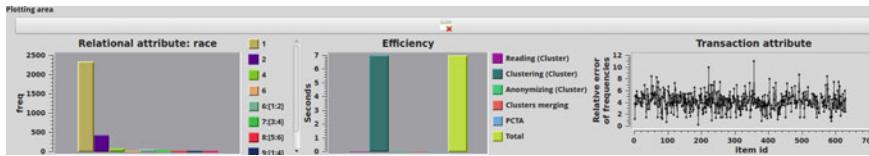


Fig. 5.15 The plotting area

5.4.6 Comparing Different Methods

Typically, users want to evaluate more than one methods and select the most beneficial. In this case, the Methods Comparison interface of SECRETA can be used (by clicking on the button). This interface enables the user to:

- Select an algorithm for anonymizing relational attributes (e.g., demographics).
- Select an algorithm for anonymizing diagnosis codes.
- Select an algorithm that will be used as a bounding method.
- Set the values for parameters that will be fixed, as described above.
- Choose one varying parameter, along with its start/end value and step.

The above actions define a *configuration*, which is added into the experimenter area on the top-right pane (Fig. 5.16) by pressing the button. Similarly, the user may create additional configurations. After the methods are applied, the user may select various graphs, which will be displayed in the plotting area.

Experimental configurations							
	Rel Method	Trans Method	Bound Method	k	m	d	Priv constr
1	Cluster	Apriori	R_Merge_R	25	2	0.4	-
2	Cluster	VPA	R_Merge_R	25	2	0.4	-
3	Cluster	PCTA	R_Merge_R	25	2	0.4	1 2 101 102 96 15

Fig. 5.16 The configurations editor

5.5 Conclusion and Future Work

In this chapter, we provided an overview of the SECRETA system, which allows applying and evaluating different anonymization algorithms on data with relational and/or transaction attributes. The SECRETA system does not require expertise in data anonymization in order to be used, and has several useful features for evaluating different algorithms and reporting their achieved level of data utility.

SECRETA can be a useful tool for medical data owners, as it enables them to effortlessly evaluate a range of anonymization methods to select the most appropriate one for a specific data publishing need. However, SECRETA is a tool that can be extended and improved along many dimensions, including the following:

- Inclusion of additional data anonymization algorithms. For example, we plan to incorporate algorithms that enforce l -diversity [12], t -closeness [10], and ρ -uncertainty [1].
- Offering of SECRETA's functionality through an API that developers can use to call anonymization methods or evaluation tools from their own software.
- Provisioning of a command line execution mode, for faster data anonymization of larger datasets by avoiding the overhead of the GUI.
- Offering of SECRETA's functionality as a web service. Currently, a user has to download the software and install it in a linux environment. On the contrary, a web-based version of SECRETA could be executed on any operating system with Internet access. Also, new releases of the software would be directly available to end users, without any actions from their side.

References

1. Cao, J., Karras, P., Raïssi, C., Tan, K.: ρ -uncertainty: inference-proof transaction anonymization. *VLDB* **3**(1), 1033–1044 (2010)
2. Dai, C., Ghinita, G., Bertino, E., Byun, J.W., Li, N.: TIAMAT: a tool for interactive analysis of microdata anonymization techniques. *VLDB* **2**(2), 1618–1621 (2009)
3. Dwork, C.: Differential privacy. In: ICALP, pp. 1–12 (2006)
4. Fung, B.C.M., Wang, K., Yu, P.: Top-down specialization for information and privacy preservation. In: 21st International Conference on Data Engineering (ICDE), pp. 205–216 (2005)
5. Gkoulalas-Divanis, A., Loukides, G.: PCTA: privacy-constrained clustering-based transaction data anonymization. In: 2011 International Workshop on Privacy and Anonymity in Information Society (PAIS), pp. 1–10 (2011)
6. Gkoulalas-Divanis, A., Loukides, G.: Anonymization of Electronic Medical Records to Support Clinical Analysis, 1st edn. Springer, New York (2013)
7. Gkoulalas-Divanis, A., Loukides, G., Sun, J.: Publishing data from electronic health records while preserving privacy: a survey of algorithms. *J. Biomed. Inform.* **50**, 4–19 (2014). doi: [10.1016/j.jbi.2014.06.002](https://doi.org/10.1016/j.jbi.2014.06.002). <http://dx.doi.org/10.1016/j.jbi.2014.06.002>
8. LeFevre, K., DeWitt, D., Ramakrishnan, R.: Incognito: efficient full-domain k -anonymity. In: SIGMOD, pp. 49–60 (2005)
9. LeFevre, K., DeWitt, D., Ramakrishnan, R.: Mondrian multidimensional k -anonymity. In: ICDE, p. 25 (2006)
10. Li, N., Li, T., Venkatasubramanian, S.: t -closeness: privacy beyond k -anonymity and l -diversity. In: ICDE, pp. 106–115 (2007)
11. Loukides, G., Gkoulalas-Divanis, A., Malin, B.: COAT: constraint-based anonymization of transactions. *Knowl. Inf. Syst.* **28**(2), 251–282 (2011)
12. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkatasubramaniam, M.: l -diversity: privacy beyond k -anonymity. In: ICDE, p. 24 (2006)
13. Nergiz, M.E., Atzori, M., Clifton, C.: Hiding the presence of individuals from shared databases. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, SIGMOD '07, pp. 665–676. ACM, New York (2007). doi: [10.1145/1247480.1247554](https://doi.org/10.1145/1247480.1247554). <http://doi.acm.org/10.1145/1247480.1247554>
14. Poulis, G., Loukides, G., Gkoulalas-Divanis, A., Skiadopoulos, S.: Anonymizing data with relational and transaction attributes. In: ECML/PKDD, pp. 353–369 (2013)
15. Poulis, G., Skiadopoulos, S., Loukides, G., Gkoulalas-Divanis, A.: Distance-based k^m -anonymization of trajectory data. In: 2013 IEEE 14th International Conference on Mobile Data Management, vol. 2, pp. 57–62, 3–6 June 2013, Milan (2013). doi: [10.1109/MDM.2013.66](https://doi.org/10.1109/MDM.2013.66). <http://dx.doi.org/10.1109/MDM.2013.66>
16. Poulis, G., Skiadopoulos, S., Loukides, G., Gkoulalas-Divanis, A.: Select-organize-anonymize: a framework for trajectory data anonymization. In: 13th IEEE International Conference on Data Mining Workshops, ICDM Workshops, pp. 867–874, 7–10 December 2013, TX (2013). doi: [10.1109/ICDMW.2013.136](https://doi.org/10.1109/ICDMW.2013.136). <http://dx.doi.org/10.1109/ICDMW.2013.136>
17. Poulis, G., Gkoulalas-Divanis, A., Loukides, G., Skiadopoulos, S., Tryfonopoulos, C.: SECRETA: a system for evaluating and comparing relational and transaction anonymization algorithms. In: Proceedings of the 17th International Conference on Extending Database Technology, EDBT 2014, pp. 620–623, 24–28 March 2014, Athens (2014). doi: [10.5441/002/edbt.2014.58](https://doi.org/10.5441/002/edbt.2014.58). <http://dx.doi.org/10.5441/002/edbt.2014.58>
18. Poulis, G., Skiadopoulos, S., Loukides, G., Gkoulalas-Divanis, A.: Apriori-based algorithms for k^m -anonymizing trajectory data. *Trans. Data Privacy* **7**(2), 165–194 (2014). <http://www.tdp.cat/issues11/abs.a194a14.php>
19. Prasser, F., Kohlmayer, F., Kuhn, K.A.: Arx-a comprehensive tool for anonymizing biomedical data. In: AMIA Annual Symposium (2014)

20. Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Uncertainty Fuzziness Knowl.-Based Syst.* **10**(5), 557–570 (2002)
21. Terrovitis, M., Mamoulis, N., Kalnis, P.: Privacy-preserving anonymization of set-valued data. *PVLDB* **1**(1), 115–125 (2008)
22. Terrovitis, M., Mamoulis, N., Kalnis, P.: Local and global recoding methods for anonymizing set-valued data. *VLDB J.* **20**(1), 83–106 (2011)
23. Xiao, X., Wang, G., Gehrke, J.: Interactive anonymization of sensitive data. In: SIGMOD, pp. 1051–1054 (2009)
24. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Chee Fu, A.W.: Utility-based anonymization using local recoding. In: SIGKDD, pp. 785–790 (2006)

Chapter 6

Putting Statistical Disclosure Control into Practice: The ARX Data Anonymization Tool

Fabian Prasser and Florian Kohlmayer

Abstract The sharing of sensitive personal data has become a core element of biomedical research. To protect privacy, a broad spectrum of techniques must be implemented, including data anonymization. In this article, we present ARX, an anonymization tool for structured data which supports a broad spectrum of methods for statistical disclosure control by providing (1) models for analyzing re-identification risks, (2) risk-based anonymization, (3) syntactic privacy criteria, such as k -anonymity, ℓ -diversity, t -closeness and δ -presence, (4) methods for automated and manual evaluation of data utility, and (5) an intuitive coding model using generalization, suppression and microaggregation. ARX is highly scalable and allows for anonymizing datasets with several millions of records on commodity hardware. Moreover, it offers a comprehensive graphical user interface with wizards and visualizations that guide users through different aspects of the anonymization process. ARX is not just a toolbox, but a fully-fledged application, meaning that all implemented methods have been harmonized and integrated with each other. It is well understood that balancing privacy and data utility requires user feedback. To facilitate this interaction, ARX is highly configurable and provides various methods for exploring the solution space.

6.1 Introduction

Collaborative collection and sharing of sensitive personal data have become indispensable for biomedical research. To secure privacy in complex research environments, a broad spectrum of measures must be implemented, including legal, contractual as well as technical measures. Data anonymization is a central building block in this context. It aims at sanitizing datasets in ways that prevent attackers from breaching the subjects' privacy. Different definitions of privacy and techniques

The authors “Fabian Prasser” and “Florian Kohlmayer” contributed equally to this work.

F. Prasser (✉) • F. Kohlmayer

Chair for Biomedical Informatics, Technische Universität München, München, Germany

e-mail: prasser@in.tum.de; florian.kohlmayer@tum.de

for sanitizing datasets have been proposed [13, 15, 20]. As sanitization inevitably leads to information loss and, thus, a decrease in data quality, the balance between privacy protection and utility for a specific use case has to be considered.

In this work, we will focus on statistical disclosure control with syntactic privacy models [15, 20]. We will relate our work to semantic privacy models [13] in Sects. 6.5.1 and 6.5.2.

6.1.1 Background

Methods of statistical disclosure control offer the dynamic means needed to balance privacy and utility. The spectrum of methods is broad [8, 17, 20]. Syntactic privacy criteria are used to express privacy requirements. Coding models, often generalization and suppression of attribute values, are utilized to transform datasets in such a way that they adhere to the specified privacy requirements. Risk models are used to estimate risks of re-identification, which are an inherent aspect of many privacy models. Utility measures allow estimating the suitability of the resulting datasets for specific usage scenarios. Methods of statistical disclosure control have been recommended by official working groups [1], as they can effectively protect sensitive biomedical data while resulting in a higher data utility compared to simple heuristics such as the HIPAA Safe Harbor method [36]. Still, however, they are used rarely. An important reason for this is that effective statistical disclosure control requires the use of a large set of techniques in an integrated manner.

Firstly, a balancing of privacy and utility can be achieved by using different methods, as well as by varying their respective parameters. Secondly, data may be shared with different users under differing laws and regulations, leading to different privacy requirements. Thirdly, different usage scenarios only allow for different types of transformations to be applied to data. Finally, different techniques have been developed for anonymizing different types of data. To cope with this broad spectrum of requirements, systems need to support various privacy models, risk models, methods for evaluating data utility and types of data transformations. As each of these individual methods is already complex by itself, it is not surprising that neither best practices nor concise guidelines are available to assist data controllers in dealing with these complexities. Moreover, tools that provide adequate access to the variety of methods needed are still in their infancy.

To improve upon existing tools, methods need not only be implemented but also integrated with each other to enable exploiting their full potential. This is non-trivial. For example, it requires a fair amount of effort to make sure that privacy models can be combined with each other, with different methods for automatic evaluation of data utility, and with different transformation models. It is cumbersome to adjust the details and to make sure that all parameters and results can be customized and visualized in the user interface. Such development efforts are usually not undertaken when developing research prototypes. Overall, this imposes a high barrier to data

controllers and researchers who want to test the practical usefulness of methods for data anonymization, develop guidelines and ultimately implement them into day-to-day data processing.

6.1.2 *Objectives and Outline*

To overcome this gap, we have put extensive efforts in developing ARX, a data anonymization tool which, to our knowledge, provides the most comprehensive support of methods for statistical disclosure control to date. Moreover, ARX is not just a toolbox but a fully-fledged application, meaning that all implemented methods have been harmonized and integrated with each other. It is well understood that balancing privacy and data utility may often require user feedback. For this reason, ARX is highly configurable and provides various methods for exploring the solution space. While ARX can be seen as a domain-independent tool, it has been designed with a specific focus on the biomedical domain. Its highlights include:

Risk analyses: ARX implements multiple methods for estimating re-identification risks, including two different super-population models [11].

Risk-based anonymization: ARX provides privacy criteria that utilize the implemented risk models. It can automatically transform datasets to ensure that risks are below a user-defined threshold.

Syntactic privacy criteria: ARX implements a variety of further syntactic privacy criteria, including k -anonymity [48], three variants of ℓ -diversity [32, 35], two variants of t -closeness [32] and δ -presence [40]. Moreover, ARX supports arbitrary combinations of privacy criteria (including risk-based methods).

Utility evaluation: ARX implements several methods for evaluating data utility, manually as well as automatically [3, 16, 24, 30]. Methods for automated utility evaluation are highly configurable.

Intuitive transformation model: ARX uses global recoding via full-domain generalization combined with local recoding via tuple suppression [27]. This coding scheme has been recommended for the biomedical domain [16]. Additionally, it supports top- and bottom-coding, as well as microaggregation [8].

Optimality: ARX implements a globally-optimal search strategy that will automatically classify the solution space and determine the transformation that is optimal according to the defined utility measure [26].

Scalability: ARX has been designed from the ground up to be highly scalable, and it is able to handle very large datasets (with millions of data entries) on commodity hardware [45].

Compatibility: ARX comes with facilities for importing data from common sources, such as CSV files, spreadsheet programs (MS Excel) and relational database systems (e.g., MS SQLServer, MySQL, PostgreSQL) [44].

Comprehensive GUI: ARX provides a fully-fledged graphical interface with features such as wizards for creating generalization hierarchies and various visualizations of data utility, risk estimates and the solution space [44].

Carefully designed API: ARX also offers an application programming interface (API) that allows accessing all of its features. All methods are first-class citizens in the API, as well as the GUI [44].

Cross-platform: ARX is implemented in Java and available for all common operating systems. The ARX user interface is built with the Standard Widget Toolkit (SWT) which provides a native look and feel on all platforms.

Open source: ARX is open source software, fostering reviews by the community and allowing users to tailor it to their requirements [2].

Implementing an integrated application with support for such a broad spectrum of anonymization methods is non-trivial. Many methods are inherently computationally complex and thus require the development of sophisticated data structures and algorithms to implement them in an efficient manner. Moreover, implementations need to be designed carefully, to ensure that they are inter-operable with the many other methods relevant to data anonymization.

In this chapter, we present ARX and sketch some of its most important design and implementation aspects. The remainder of the chapter is structured as follows. In Sect. 6.2 we give an in-depth overview of the ARX system. We present the system architecture and introduce the application programming interface as well as the graphical user interface. In Sect. 6.3 we describe selected implementation details. In Sect. 6.4 we present a brief experimental evaluation, focusing on the more advanced methods supported by ARX. Last, we discuss our work and conclude this chapter.

6.2 The ARX Data Anonymization Tool

There are three different types of data anonymization typically performed with personal health data:

1. Masking identifiers in unstructured data, e.g., removing HIPAA identifiers [53] from clinical notes using machine learning approaches or regular expressions [25].
2. Privacy-preserving data analysis in an interactive scenario. Methods from this area include interactive *differential privacy* [13] and *query-set-size control*.
3. Anonymization of static structured microdata, where data is expected to be in a tabular form with each row corresponding to the data about one individual [20].

The focus of our work, ARX, is on the anonymization of static structured microdata, which may, for example, be used to de-identify data that has been collected for biomedical research prior to sharing it with others.

6.2.1 Background

When anonymizing structured data, the general attack vector assumed is *linkage* of a sensitive dataset with an identified dataset (or similar background knowledge about individuals). The attributes that may be used for linkage are termed *quasi-identifiers* (or indirect identifiers, or keys). Such attributes are not identifiers per-se but may in combination be used for linkage. Moreover, it is assumed that they cannot simply be removed from the dataset as they may be required for analyses and that they are likely to be available to an attacker.

Furthermore, it is assumed that *directly identifying* information (such as names) have already been removed from the dataset. An example dataset with different types of attributes is shown in Fig. 6.1. Three types of privacy threats are commonly considered [34]:

- Membership disclosure** means that data linkage allows an attacker to determine whether or not data about an individual is contained in a dataset [40]. While this does not directly disclose any information from the dataset itself, it may allow an attacker to infer meta-information. In our example, the data is from a cancer registry and it can thus be inferred that an individual has or has had cancer. While this deals with implicit sensitive attributes (meaning attributes of an individual that are not contained in the dataset), other disclosure models deal with explicit sensitive attributes.
- Attribute disclosure** may be achieved even without linking an individual to a specific item in a dataset [35]. It protects *sensitive attributes*, which are attributes from the dataset with which individuals are not willing to be linked with. As such, they might be of interest to an attacker and, if disclosed, could cause harm to data subjects. As an example, linkage to a set of data entries allows inferring information if all items share a certain sensitive attribute value (breast cancer in our example).
- Identity disclosure** (or *re-identification*) means that an individual can be linked to a specific data entry [50]. This is a very serious type of attack, as it has legal consequences for data owners according to many laws and regulations worldwide. From the definition it also follows that an attacker can learn all sensitive information contained in the data entry about the individual.

Fig. 6.1 Example cancer dataset: types of attributes and types of disclosure

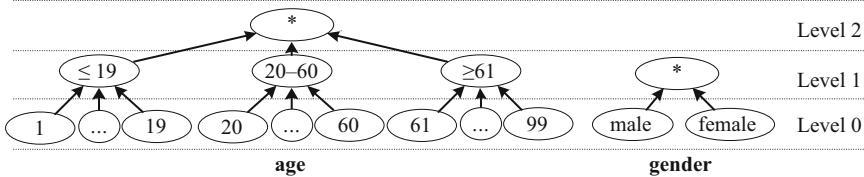


Fig. 6.2 Generalization hierarchies for attributes age and gender

To counter privacy threats with ARX, data is transformed with *generalization hierarchies*. The tool supports hierarchies for both categorical and continuous attributes. Examples are shown in Fig. 6.2. Here, values of the attribute *age* are transformed into age groups, while values of the attribute *gender* can only be suppressed. Generalization hierarchies are well suited for categorical attributes. They can also be used for continuous attributes by performing on-the-fly categorization. In ARX, this categorization is implemented by representing generalization strategies in a functional manner, e.g., as a set of intervals.

To increase data utility, attribute generalization is combined with the *suppression* of data records. This means that rows which violate the privacy criteria (i.e., *outliers*) are automatically removed from the dataset, while the total number of removed records is kept under a specified threshold, which is called the *suppression limit*. As a result, less generalization is required to ensure that the remaining records satisfy the privacy model.

ARX implements multi-dimensional global recoding with full-domain generalization and local recoding with tuple suppression. This combination of methods has been found to be well-suited for the biomedical domain, as it is (1) intuitive, (2) produces datasets that are well suited for analyses by epidemiologists and, (3) can be configured by non-experts [16]. Configuration (and thus balancing of privacy and utility) may be performed by altering generalization hierarchies or by choosing a suitable transformation from the solution space. For continuous variables, ARX also supports *microaggregation* [8]. This means that the values of a quasi-identifier within one group are transformed by applying aggregate functions, such as the arithmetic or geometric mean. In the anonymized table, the entire group of values is then replaced with the result of the aggregate function. In contrast to generalization or suppression, this method can be used to preserve the data type and scale of measure of a numerical variable.

When using global recoding with full-domain generalization, the search space can be modeled as a *generalization lattice*, which is a partially ordered set of all possible combinations of generalization levels of each attribute. Lattices can be visualized with Hasse diagrams, which in our context means to draw the transitive reduction of a set of transformations in an acyclic graph, where each node is connected to all of its direct successors and predecessors [55]. Each node represents one transformation and defines generalization levels for all quasi-identifiers. An arrow denotes that a transformation is a direct generalization of a more specialized transformation, which can be created by incrementing one of the generalization

Fig. 6.3 Example search space

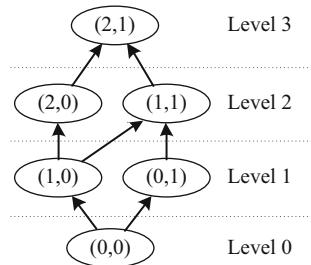


Table 6.1 Example dataset and the result of applying the transformation (1,0)

Quasi-identifying		Insensitive	Sensitive
Age	Gender	State	Diagnosis
34	Male	NY	Pneumonia
45	Female	MS	Gastritis
66	Male	NY	Gastritis
70	Male	TX	Pneumonia
35	Female	AL	Pneumonia
21	Male	AL	Gastritis
18	Female	TX	Pneumonia
19	Female	MS	Gastritis

Quasi-identifying		Insensitive	Sensitive
Age	Gender	State	Diagnosis
20–60	Male	NY	Pneumonia
20–60	Female	MS	Gastritis
≥61	Male	NY	Gastritis
≥61	Male	TX	Pneumonia
20–60	Female	AL	Pneumonia
20–60	Male	AL	Gastritis
≤19	Female	TX	Pneumonia
≤19	Female	MS	Gastritis

levels defined by its predecessor. An example utilizing the hierarchies from Fig. 6.2 is shown in Fig. 6.3. The original dataset is at the bottom (0, 0), whereas the transformation with maximal generalization (2, 1) is at the top.

Table 6.1 shows an example dataset in which the attributes `age` and `gender` are considered quasi-identifying, `state` is considered insensitive, and `diagnosis` is considered to be a sensitive attribute. Moreover, the table shows the result of applying the transformation (1, 0) to the quasi-identifiers, which leaves `gender` as-is and generalizes `age` to the first level of the associated hierarchy.

6.2.2 Overview

In this section, we will present a short overview of the basic functionalities of ARX in terms of supported privacy models, risk models and methods for utility evaluation.

6.2.2.1 Privacy Models

To prevent privacy breaches, ARX implements privacy models that apply syntactic privacy criteria on a dataset. In these privacy models, assumptions are made about potential attackers and (likely) background knowledge, as well as goals, by

classifying attributes according to the above definitions. The tool supports arbitrary combinations of the following privacy models:

- ***k*-Anonymity** is the best known privacy model [50]. It aims at protecting datasets from identity disclosure. A dataset is *k-anonymous* if, regarding the quasi-identifiers, each data item cannot be distinguished from at least $k-1$ other data items. This property can be used to define *equivalence classes* of indistinguishable entries [48]. The output dataset from Table 6.1 fulfills 2-anonymity.
- **ℓ -Diversity** aims at protecting datasets against attribute disclosure [35]. It requires that each sensitive attribute must have at least ℓ “well represented” values in each equivalence class. ARX implements three variants, which use different notions of diversity: *distinct- ℓ -diversity*, which requires ℓ different sensitive values per class [32], as well as *recursive-(c, ℓ)-diversity* and *entropy- ℓ -diversity* [35]. The output dataset from Table 6.1 fulfills distinct-2-diversity.
- **t -Closeness** overcomes some limitations of ℓ -diversity but it may also be more difficult to achieve [32]. It requires that the distance between the distribution of sensitive values in each equivalence class and their overall distribution in the dataset must be lower than a given threshold [32]. ARX implements two variants, one of which uses generalization hierarchies to compute distances between different distributions of an attribute.
- **δ -Presence** aims at protecting datasets against membership disclosure [40]. It requires that the disclosed dataset is explicitly modeled as a subset of a larger dataset that represents the attacker’s background knowledge. The criterion enforces restrictions on the probabilities with which it can be determined whether or not an individual from the global dataset is contained in the research subset. Upper bounds for these probabilities are calculated based on the sizes of equivalence classes [40].
- **Risk-based anonymization** can be used to automatically transform datasets to ensure that estimated re-identification risks fall below a given threshold [11].

6.2.2.2 Risk Analysis and Risk-Based Anonymization

In ARX, *super-population models* may be used to estimate *population uniqueness*, i.e., the fraction of entries in the dataset that are unique within the overall population. These statistical methods estimate characteristics of the overall population with probability distributions that are parameterized with sample characteristics. ARX provides default settings for populations, such as the USA, UK, France or Germany, and implements The sentence is misleading as the model by Zayatz is not a super-population model. The model by Hoshino, which is based on Pitman’s sampling formula [22], and a model using a slide negative binomial (SNB) distribution [6, 47], in which it is assumed that the sample has been created through random binomial sampling. Moreover, the tool implements the risk model by Zayatz [21, 61], which assumes no specific distribution of the data. Three models have also been combined into a *decision rule for biomedical datasets* which has been validated with real-world data [11].

In the tool, risk estimates can also be used for *risk-based anonymization*. The static k -anonymity model basically defines an upper bound on the re-identification risk, which is (over-) estimated with sample frequencies. ARX supports multiple relaxed variants of this criterion that utilize the aforementioned risk models to ensure that risk estimates fall below a given threshold.

6.2.2.3 Utility Evaluation

For automatically measuring data utility, ARX supports methods based on equivalence classes, which are called *single-dimensional utility metrics* in the tool. Examples include *Discernibility* [3] and *Average Equivalence Class Size* [30]. Both measures estimate information content based on the size of equivalence classes, while discernibility also involves a penalty for suppressed tuples.

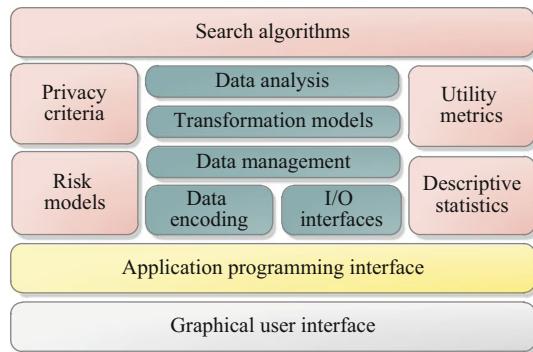
Additionally, ARX supports methods based on the actual data values. As these methods compute independent values for each attribute, which are then compiled into a global value, we call them *multi-dimensional utility metrics*. ARX supports several aggregate functions for compiling global utility measures, including *sum*, *arithmetic mean* and *geometric mean*.

Moreover, multi-dimensional metrics may be weighted, meaning that weights can be assigned to attributes to model their importance. The implemented measures include *Non-Uniform Entropy* [16], which is an information theoretic method, and *Loss* [24], which measures the extent to which the domain of an attribute is covered by its generalization. The latter method makes use of *functional representations* of generalization hierarchies (cf. Sect. 6.1.1). Our globally-optimal search strategy, *Flash*, is able to automatically determine a transformation that maximizes data utility according to the implemented utility measures [26].

6.2.2.4 Additional Features

ARX is compatible with a wide range of data processing tools. It currently features interfaces for importing data from character-separated values (CSV) files, MS Excel spreadsheets and relational database management systems (RDBMSs), such as MS SQLServer, PostgreSQL and MySQL. The syntax of CSV encoded files is detected automatically. On a semantic level, ARX supports different data types and scales of measure. Formats of data types are also detected automatically. ARX supports data cleansing during import, meaning that invalid data are replaced with a value that represents missing data. ARX safely handles missing values by making sure that different missing values match each other [7].

Fig. 6.4 High-level architecture of the ARX system



6.2.3 System Architecture

The high-level system architecture of ARX is shown in Fig. 6.4. Among the design objectives was to prevent tight coupling of subsystems and to ensure extensibility. The core modules, which normally need not be modified, provide the basis for our software. Implementation details about this module will be given in Sect. 6.3.

The I/O module provides methods for reading data from and writing data to external storage, while the data encoding module transforms the data into the format and memory layout required by the framework. The data management module deals with this internal representation and implements several optimizations (see Sect. 6.3). The transformation model is implemented on top of this representation and supports generalizing, suppressing and aggregating values. The data analysis module provides methods for grouping data records and for computing frequency distributions of attribute values. These distributions can be used to protect sensitive attributes and to apply microaggregation. The analysis module also acts as a bridge to privacy models and data utility measures.

Extensible modules are built on top of the core modules and provide implementations of privacy criteria, data utility measures, risk models, methods of descriptive statistics, and anonymization algorithms. These modules are only loosely coupled to the internals of the framework. Currently, our tool features several variants of the *Flash* algorithm but the framework can be used to implement a large set of methods [28, 44]. The application programming interface (API) is based on both the extensible and the core modules. It is also used by the graphical interface, which is thus decoupled from the internals of the system.

Figure 6.5 provides a more detailed overview of the basic design of ARX's core. It shows a simplified Unified Modeling Language (UML) class diagram, in which classes that are part of the public Application Programming Interface (API) of ARX are drawn with a thick border. The public API will be presented in more detail in the next paragraph. In ARX, the attributes of a dataset are partitioned into different **Buffers**, depending on which types of transformation (generalization or aggregation) and analyses (computation of frequency distributions) are to be performed with

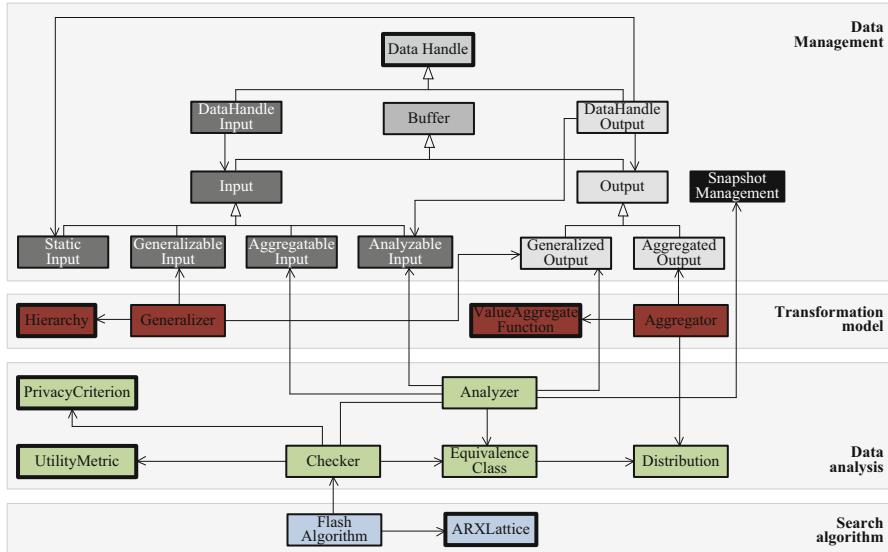


Fig. 6.5 Overview of the most important classes in ARX’s core

them. For users of the API, **DataHandles** re-integrate these separate buffers and provide transparent access. For parts of the data that are modified, separate output buffers exist that hold transformed attribute values. The transformation module reads data from the input buffers and the frequency distributions of attribute values computed by the analysis module and writes generalized (via **Generalizer**) or aggregated (via **Aggregator**) values to the output buffers.

The **Analyzer** from the data analysis module accesses data from the input and output buffers, and computes equivalence classes and frequency distributions of attribute values. It makes use of the snapshot management component from the data management module for optimizations (see Sect. 6.3). The class **Checker** acts as the main interface to anonymization algorithms. It orchestrates the process of transforming data, analyzing data and evaluating privacy criteria as well as data utility measures. Our search algorithm, *Flash*, traverses a given generalization lattice and uses this interface for evaluating properties of data transformations.

A UML diagram of the most important classes of the API of ARX is shown in Fig. 6.6. The API has packages for (1) data import and configuration, (2) hierarchy creation, (3) privacy criteria specification, (4) data utility quantification, (5) utility evaluation, (6) representation of the solution space, and (7) risk analyses.

For hierarchy creation, ARX provides several **ValueAggregateFunctions**, which can also be used for microaggregation. Privacy criteria are divided into class-based and sample-based. The latter are evaluated for the complete dataset instead of individual equivalence classes. Currently, all instances of **RiskBasedPrivacyCriterion** are sample-based. These criteria make use of risk estimators provided by the corresponding package. Risk estimators are

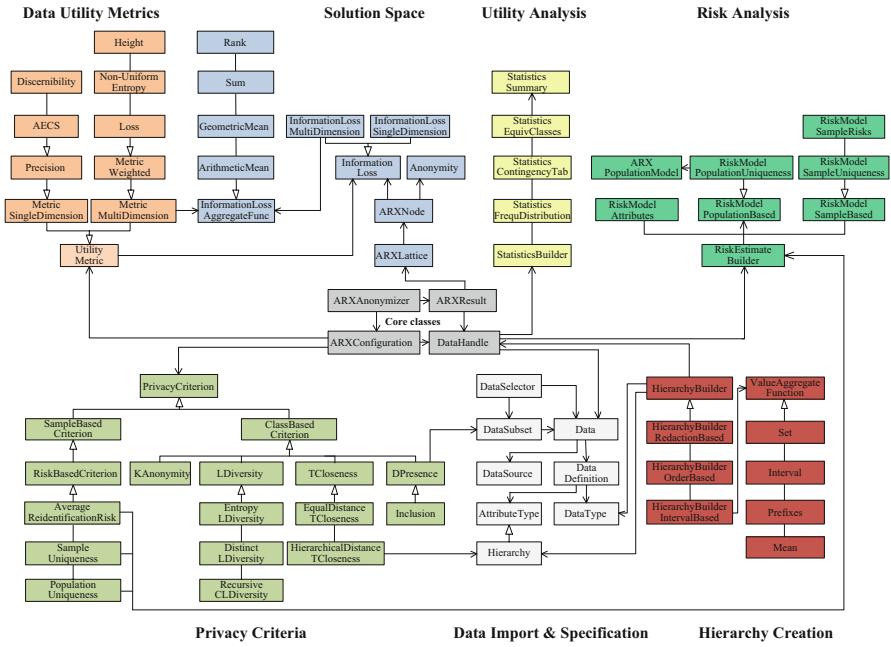


Fig. 6.6 Overview of the most important classes in ARX’s application programming interface

implemented for analyzing (groups of) attributes and for analyzing the dataset based on sample characteristics or super-population models.

Data utility measures are divided into single-dimensional and multi-dimensional **UtilityMetrics** (cf. Sect. 6.2.2.3). The utility analysis package provides basic methods of descriptive statistics for evaluating input and output data. It is accessible via **DataHandles**.

The classes **ARXConfiguration**, **ARXAnonymizer** and **ARXResult** act as the main interfaces to ARX’s functionalities. Data and attribute types are provided as static variables from the **DataType** and **AttributeType** classes, respectively. The class **DataHandle** allows users to interact with the data (read-only), by performing operations such as sorting, swapping rows or accessing cell values. Handles can be obtained for input data, output data and horizontal subsets of such data. Data handles, representing input and derived output data, are linked with each other, meaning that operations performed on one representation are transparently performed on all other representations as well. For example, sorting the input data sorts the output data analogously. The class **ARXLattice** offers several methods for exploring the solution space and for obtaining information about the properties of transformations, represented by instances of the class **ARXNode**.

All features that are accessible via the graphical interface are also accessible via the public API. However, the aim of the programming interface is to provide de-identification methods to other software systems and we note that interaction

with the software library will often be simpler than interaction with the graphical tool. Programmatic access will usually rely on ARX's ability to automatically determine a solution to a privacy problem.

6.2.4 Application Programming Interface

In this section, we will present a running example for anonymizing a dataset with the API. While this example naturally does not make use of all of ARX's features, it covers many of the more advanced functionalities of the software. This includes (1) importing data from a relational database system, (2) using functional representations of generalization hierarchies, (3) enforcing t -closeness using generalization hierarchies for computing distances, (4) performing a risk analysis that uses super-population models and Dankar et al.'s decision rule for biomedical datasets, as well as, (5) automatically determining an optimal solution with regards to a utility measure that is parameterized with attribute weights and preferences about how the coding model should be applied. We assume a dataset with three quasi-identifiers, `age`, `gender` and `zip code` as well as one sensitive attribute, `LDL cholesterol`.

We start by creating generalization hierarchies for the four attributes. ARX implements methods for creating generalization hierarchies by mapping values into intervals, by ordering and grouping values, and by masking characters. Listing 6.1 shows an example of how to use intervals for categorizing the attribute `LDL cholesterol` into six categories ranging from *very low* (`LDL cholesterol` lower than 1.8) to *very high* (`LDL cholesterol` greater than 4.9). On the next level of the hierarchy, the categories *very low* and *low* are merged into the category *low*, the categories *normal* and *borderline high* are merged into the category *normal*, and the categories *high* and *very high* are merged into the category *high*. On the next level, *low* and *normal* are merged into *low-normal*, while *high* remains unchanged. Such functional representations of hierarchies enable ARX to handle continuous variables via categorization.

```

1 // Create builder that uses intervals for categorization
2 HierarchyBuilderIntervalBased<Double> bLDL = HierarchyBuilderIntervalBased
3                                         .create(DataType.DECIMAL);
4
5 // Define base intervals (level 1)
6 bLDL.addInterval(0.0d, 1.8d, "very low");
7 bLDL.addInterval(1.8d, 2.6d, "low");
8 bLDL.addInterval(2.6d, 3.4d, "normal");
9 bLDL.addInterval(3.4d, 4.1d, "borderline high");
10 bLDL.addInterval(4.1d, 4.9d, "high");
11 bLDL.addInterval(4.9d, 10d, "very high");
12
13 // Define groups (levels 2 and 3)
14 bLDL.getLevel(0).addGroup(2, "low").addGroup(2, "norm").addGroup(2, "high");
15 bLDL.getLevel(1).addGroup(2, "low-norm").addGroup(1, "high");

```

Listing 6.1 Categorizing the attribute `LDL cholesterol` and creating a generalization hierarchy

In Listing 6.2 we use the same type of functional hierarchy for the discrete attribute `age`. In this case, we define one single base interval for the range [0, 5]. This interval will automatically be repeated within the defined overall range [0, 80]. On subsequent levels, we merge instances of this interval, two at a time, resulting in intervals of sizes 10, 20, 40 and 80 on the higher levels. Values that are lower than the lower bound of the range (0) will raise an error, while values that are larger than the upper bound of the range (80) will be top-coded, meaning that they will be replaced with the label > 80. Labels for the automatically derived groups of values will be created with a user-defined aggregate function. In the example we use the `interval` function, which will create labels such as “[lower, upper]” for the according ranges.

```

1 // Create builder for values in range [0, 120] with top-coding for values > 80
2 HierarchyBuilderIntervalBased<Long> bAge = HierarchyBuilderIntervalBased
3         .create(DataType.INTEGER, new Range<Long>(0, 0, 0),
4             new Range<Long>(80, 80, 120));
5
6 // Use intervals as aggregate function, lower included, upper excluded
7 bAge.setAggregateFunction(DataType.INTEGER.createAggregate()
8                     .createIntervalFunction(true, false));
9
10 // Define base interval on level 1
11 bAge.addInterval(0, 5);
12
13 // Define groups on levels 2, 3, 4 and 5
14 bAge.getLevel(0).addGroup(2); // [0, 10[, [10, 20[, [20, 30[, [30, 40[, ...
15 bAge.getLevel(1).addGroup(2); // [0, 20[, [20, 40[, [40, 60[, [60, 80[
16 bAge.getLevel(2).addGroup(2); // [0, 40[, [40, 80[
17 bAge.getLevel(3).addGroup(2); // [0, 80[

```

Listing 6.2 Creating a generalization hierarchy for the attribute `age`

Next, we use character masking for creating a hierarchy for the attribute `zip code`, as is shown in Listing 6.3. In this example, all values of the attribute will be padded to the same size and left-aligned. Finally, characters will be masked from right to left, e.g., creating the following generalizations for the value 81667 → 8166★ → 816★★ → 81★★★ → 8★★★★ → ★★★★★.

```

1 // Create builder that aligns from right to left and masks from left to right
2 HierarchyBuilder<?> bZipCode = HierarchyBuilderRedactionBased.create(
3             Order.RIGHT_TO_LEFT, Order.RIGHT_TO_LEFT,
4             ' ', '*');

```

Listing 6.3 Creating a generalization hierarchy for the attribute `zip code`

Finally, as can be seen in Listing 6.4, we use ordering and grouping for creating a hierarchy for the attribute `sex`. In the example, we merge `male` and `female` into a common group. For creating labels we use the `set` function, resulting in the label `{male, female}`.

```

1 // Create builder that aligns from right to left and masks from left to right
2 HierarchyBuilderOrderBased<String> bSex = HierarchyBuilderOrderBased
3                               .create(DataType.STRING);
4
5 // Define the default aggregate function
6 bSex.setAggregateFunction(DataType.STRING.createAggregate()
7                               .createSetFunction());
8
9 // Define groups
10 bSex.getLevel(0).addGroup(2); // {male, female}

```

Listing 6.4 Creating a generalization hierarchy for the attribute `sex`

We are now ready to use these generalization hierarchies for anonymizing a dataset. In our example, we will load a table (*tbl1*) from a SQLite database. This process is sketched in Listing 6.5. We define the data types of the four attributes, which means that data cleansing will automatically be performed during data import by removing values that do not conform to the given data types. Additionally, we rename the column `ldlc` to `ldl cholesterol`.

```

1 // Load driver
2 Class.forName("org.sqlite.JDBC");
3
4 // Define source database
5 DataSource source = DataSource.createJDBCSOURCE("jdbc:sqlite:test.db", "tbl1");
6
7 // Define columns to be imported
8 source.addColumn("sex", DataType.STRING);
9 source.addColumn("age", DataType.INTEGER);
10 source.addColumn("zip code", DataType.STRING);
11 source.addColumn("ldlc", "ldl cholesterol", DataType.DECIMAL);
12
13 // Create data object
14 Data data = Data.create(source);

```

Listing 6.5 Importing data from a relational database system

Finally, we define and configure the privacy model, the method for measuring data utility and our coding model. This process is shown in Listing 6.6. Firstly, we specify that `sex`, `age` and `gender` are quasi-identifiers, by associating them with the respective hierarchies. Secondly, we specify that `ldl cholesterol` is a sensitive attribute. Thirdly, we set the suppression limit to 100 %, as ARX's utility measures will automatically balance the application of generalization and suppression to achieve optimal utility. As a privacy criterion, we specify *0.2-closeness* on the sensitive attribute, which uses the respective generalization hierarchy to compute the distance between frequency distributions.

Next, we create an instance of the *Loss* metric, which will favor generalization over suppression (a parameter of 0.0 means to only apply generalization, 1.0 means to only apply suppression). We further define that `age` is our most important attribute, followed by `sex` and `zip code`. ARX will try to reduce the amount of generalization applied to more important attributes. When everything is configured accordingly we perform the anonymization process.

```

1 // Specify attribute types and assign hierarchies
2 Data.getAttributeType("sex", bSex);
3 Data.getAttributeType("age", bAge);
4 Data.getAttributeType("zip code", bZipCode);
5 Data.getAttributeType("ldl cholesterol",
6                         AttributeType.SENSITIVE_ATTRIBUTE);
7
8 // Define privacy model
9 ARXConfiguration config = ARXConfiguration.create();
10 config.setMaxOutliers(1.0d);
11 config.addPrivacyCriterion("ldl cholesterol",
12                             new HierarchicalDistanceTCloseness(0.2d, bLDL));
13
14 // Configure coding model
15 config.setMetric(Metric.createLossMetric(0.3d));
16 config.setAttributeWeight("sex", 0.5d);
17 config.setAttributeWeight("age", 0.7d);
18 config.setAttributeWeight("zip code", 0.3d);
19
20 // Anonymize
21 DataHandle result = new ARXAnonymizer().anonymize(data, config).getOutput();

```

Listing 6.6 Enforcing t -closeness on the attribute ldl cholesterol

Listing 6.7 shows how the re-identification risks of the resulting transformed data may be analyzed. In this example, we assume that the dataset is a sample from the US population and estimate population uniqueness, i.e., the fraction of entries from the sample that are estimated to be unique within the overall population. To this end, we use Dankar et al.'s decision rule [11].

```

1 ARXPopulationModel population = ARXPopulationModel.create(Region.USA);
2 double uniqueness = result.getRiskEstimator(population)
3                                     .getPopulationBasedUniquenessRisk()
4                                     .getFractionOfUniqueTuples(
5                                         PopulationUniquenessModel.DANKAR);

```

Listing 6.7 Analyzing population uniqueness of the t -close result dataset

6.2.5 Graphical User Interface

In this section, we will first describe the workflow supported by ARX's graphical user interface. We will then present the interface in more detail and describe the views that may be used to repeat the previous example with the graphical interface.

6.2.5.1 Anonymization Process

The central challenge in data anonymization is to achieve a balance between data utility and privacy. In ARX, methods that model the different aspects of this process are combined into a multi-step workflow that allows users to iteratively adjust parameters, until the result matches their requirements. As is depicted in Fig. 6.7, the basic steps consist of (1) configuring the privacy and transformation model,

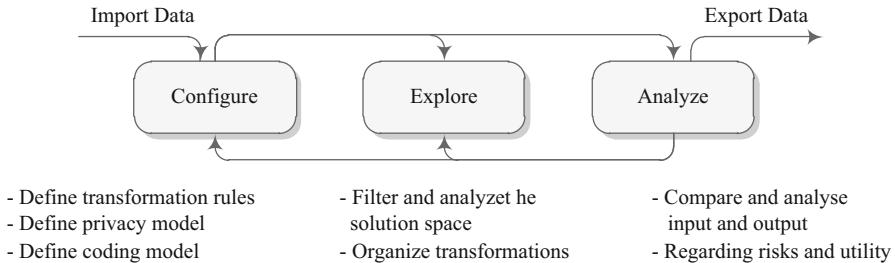


Fig. 6.7 Anonymization process implemented in ARX’s GUI

(2) exploring the solution space, and (3) analyzing input and output data. In the configuration phase, input data is loaded and annotated, generalization hierarchies are created or imported, and all further parameters such as privacy criteria, are specified.

When the solution space has been characterized by executing ARX’s anonymization algorithm, the exploration phase supports searching for privacy-preserving data transformations that fulfill the user’s requirements. To assess suitability, the analysis phase allows comparing transformed datasets to the original input dataset with methods of descriptive statistics. Moreover, risk-analyses can be performed for input data as well as its transformed representations. Based on these analyses, further solution candidates might be considered and evaluated or the configuration of the anonymization process may be altered.

6.2.5.2 Overview

The three steps from the anonymization process are mapped to four perspectives in the ARX user interface:

Configuration: In this perspective, firstly, a dataset can be imported into the tool and annotated, e.g., by characterizing attributes. Secondly, generalization hierarchies for quasi-identifiers or sensitive attributes can be generated semi-automatically by means of a wizard or imported into the tool. Thirdly, privacy criteria, the method for measuring data utility and further parameters, such as properties of the transformation model, can be specified.

Exploration: As a result of the anonymization process, a solution space is constructed and characterized based on the given parameters. This perspective allows users to browse the space of data transformations, organize and filter it according to their needs and to select further transformations for analysis.

Utility evaluation: To assess the suitability of a specific transformation for a given usage scenario, this perspective enables comparing transformations of the input dataset to the original data. To this end, it incorporates various graphical representations of results of statistical analyses and also allows for a cell-by-cell comparison.

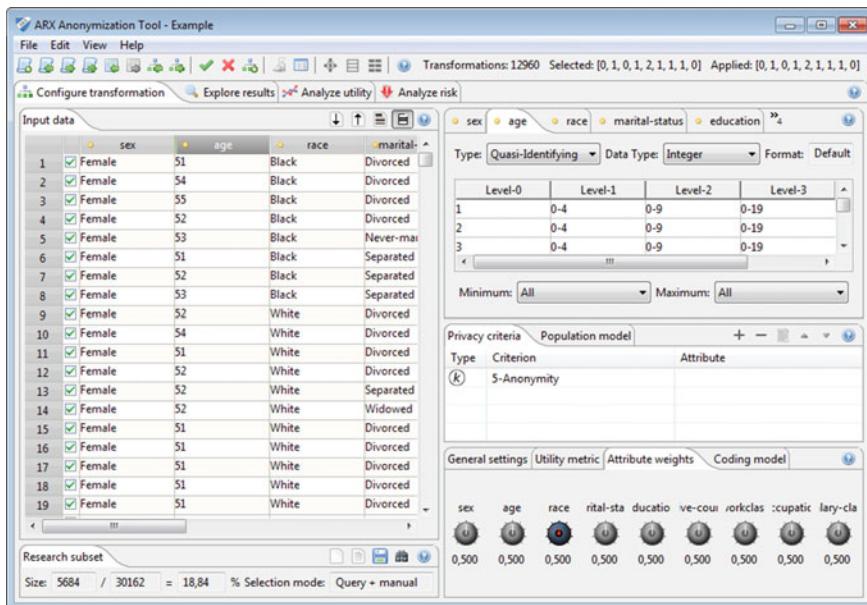


Fig. 6.8 The ARX configuration perspective

Risk analysis: In this perspective, the distribution of class sizes, the risks associated with individual quasi-identifiers, as well as sample-based and population-based risk estimates can be analyzed. The view also displays details about estimated re-identification risks obtained from different models.

6.2.5.3 Configuring the Anonymization Process

In this perspective, which is shown in Fig. 6.8, a dataset can be annotated and the anonymization process can be configured. Annotation means to classify attributes as quasi-identifying or sensitive, to specify generalization hierarchies, to define a research subset and to enter information about the overall population. Configuration means to specify the privacy model and to parameterize the transformation process. The input dataset is displayed on the left side of the view.

A querying interface, which is accessible in the lower-left corner, can be used to select a *research subset*, which is a subset of the data records that are to be included in the final anonymized dataset. This mechanism can be used for different purposes. Firstly, it can be used to enforce δ -presence on the subset, i.e., to protect the subset from membership disclosure by preventing attackers that know the overall dataset from determining whether a specific individual is or is not contained in the subset. Secondly, it can be used to manually remove data entries. The area in the upper-right corner lists all attributes and supports annotations by defining data types and

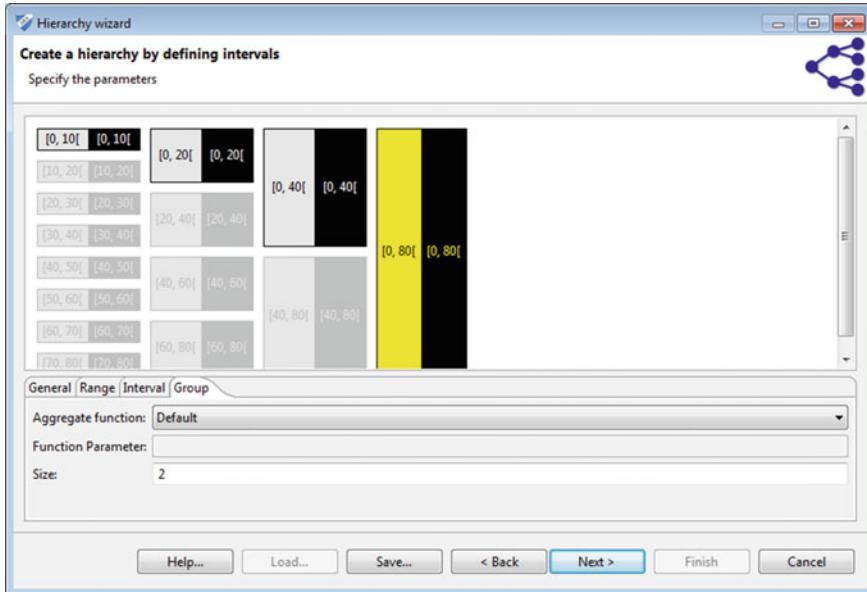


Fig. 6.9 Wizard for creating a generalization hierarchy with intervals

attribute types. The same area also shows the generalization hierarchies associated with the attributes. ARX distinguishes between four different types of attributes:

1. **Identifying attributes** will be removed from the dataset. Typical examples are names or Social Security Numbers.
2. **Quasi-identifying attributes** will be transformed by means of generalization, suppression or microaggregation. Typical examples are gender, date of birth and zip codes.
3. **Sensitive attributes** will be kept unmodified but may be subject to further constraints, such as t -closeness or ℓ -diversity. Typical examples are diagnoses.
4. **Inensitive attributes** will be kept unmodified.

For specifying generalization hierarchies, the tool offers a wizard, which supports creating hierarchies for attributes with different scales of measure (cf. Sect. 6.2.4). The wizard for creating hierarchies with intervals is shown in Fig. 6.9. Firstly, a sequence of intervals can be defined. In the next step, subsequent levels of the hierarchy can be defined, which consist of groups of intervals from the previous level. As can be seen, any sequence of intervals or groups is automatically repeated to cover the complete range of the attribute. Each element is associated with an aggregate function. These functions implement methods for creating labels for intervals, groups and values to be translated into a common generalized value.

In the configuration perspective, hierarchies are visualized as tables where each row contains the generalization rule for a single attribute value. This intuitive rep-

resentation enables compatibility with third-party applications, such as spreadsheet programs. Hierarchies can also be fine-tuned with a built-in editor that is available by right-clicking their tabular representation.

The view in the center of the right side of the perspective supports the specification of the privacy model. To this end, it shows a list of currently enabled privacy criteria and supports adding, removing and configuring criteria via according dialogs. In the second tab, population characteristics can be defined which are required for risk analyses and risk-based anonymization.

Options for configuring the transformation process are implemented in the views on the bottom of the right side of the perspective. In the general settings, the suppression limit can be specified and further performance-related parameters may be adjusted. In the second tab, utility measures can be selected and configured. The other two tabs allow users to parameterize the transformation model by weighting attributes (selected in the screenshot shown in Fig. 6.8) and by prioritizing different types of data recoding.

6.2.5.4 Exploring the Solution Space

When the search space has been characterized based on the given parameters, the exploration perspective allows users to browse potential solutions, organize them and filter them according to their requirements. The aim of the perspective (see Fig. 6.10) is to select a set of interesting transformations for further evaluation.

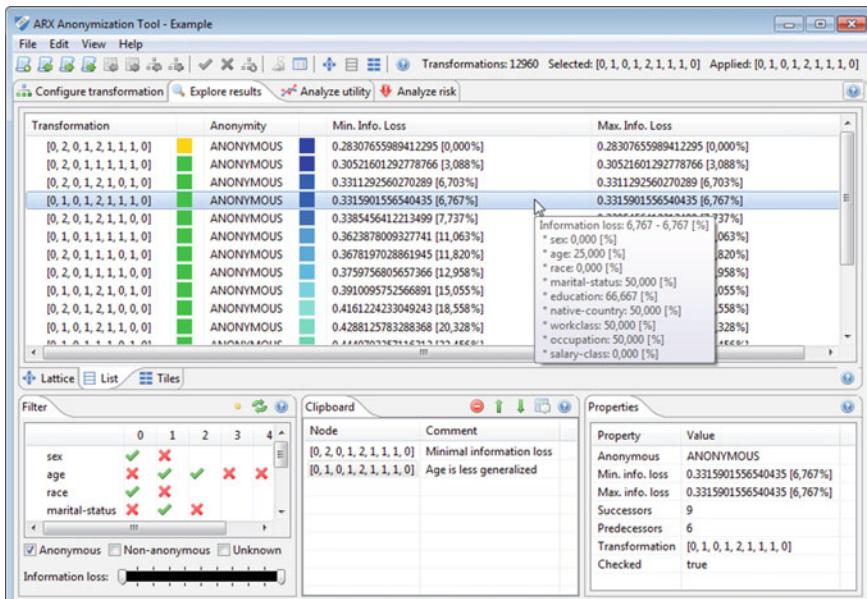


Fig. 6.10 The ARX exploration perspective

In the center of this view a subset of the solution space is visualized. In the example, transformations are presented as a list, which is ordered by data utility. Each transformation is represented by the generalization levels that it defines for the quasi-identifiers in the input dataset. Transformations are characterized by four different colors: *green* indicates that a transformation results in an anonymous dataset; *red* indicates that a transformation does not result in an anonymous dataset; *yellow* indicates that a transformation is the global optimum; and *gray* indicates that the anonymity property of a transformation is unknown as a result of pruning. If such a transformation is applied to the dataset, its actual anonymity property will be computed in the background and the state of the search space will be updated.

The lower-left area allows filtering the displayed subset of the solution space, e.g., by restricting transformations to certain generalization levels or by defining thresholds on data utility. Transformations can also be added to a clipboard, which is located in the lower-central part of the view. Here, transformations can be annotated and organized. The lower-right area displays basic information about the currently selected transformation. Right-clicking a transformation brings up a context menu, which allows to apply it to the dataset. The result can then be exported or analyzed.

Two further visualizations of the solution space are provided. Firstly, it may also be visualized as a Hasse diagram. Here, the same colors as in the list view are used to encode the anonymity property. The view is interactive and supports zooming. Secondly, the solution space may also be visualized as a set of tiles, which are ordered and colored by data utility. In contrast to the other two representations, this view is capable of displaying a large number of transformations simultaneously.

6.2.5.5 Evaluating Data Utility

This perspective enables users to compare transformed data representations to the original dataset. An example is shown in Fig. 6.11. It displays the input dataset on the left side and the output dataset on the right side. Both tables are synchronized when scrolling, supporting cell-by-cell comparisons. The data can be sorted by a single attribute, or by all quasi-identifiers, which will also highlight the resulting equivalence classes if applied to an output dataset (cf. Fig. 6.11). For comparing two data representations, the view displays typical descriptive statistics, both in tabular and graphical form. They include uni-variate and bi-variate methods, such as empirical distributions, central tendency and dispersion as well as contingency tables. The screenshot shows graphical representations of contingency tables. Together with statistical measures, this view also displays information reflecting the anonymization process. Examples include a shallow specification of the configuration and metadata about the resulting dataset, such as the number of suppressed data entries and the resulting data utility.

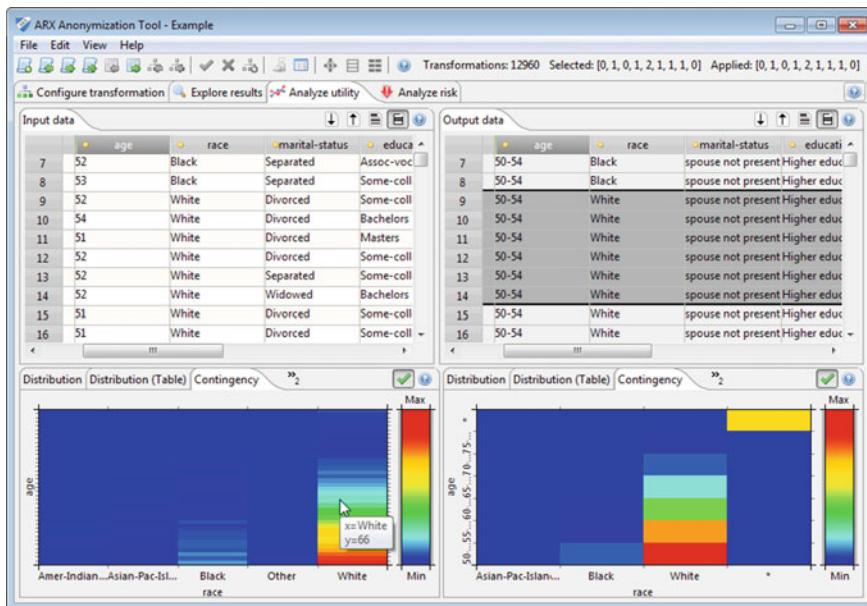


Fig. 6.11 The ARX utility evaluation perspective

6.2.5.6 Analyzing Re-identification Risks

With this perspective the re-identification risks of an input or output dataset can be analyzed with several sample-based or population-based measures. Moreover, this perspective provides an overview of the distribution of the sizes of the equivalence classes (see Sect. 6.2.2) in the dataset and supports analyzing the risks associated with groups of attributes to find quasi-identifiers.

In the screenshot shown in Fig. 6.12, the distribution of class sizes is displayed for both input and output. The view in the lower-right corner shows a comparison of the estimated population uniqueness from three risk models for different sample fractions. It also displays results of the decision rule proposed and validated for clinical datasets by Dankar et al., which selects a different risk model for different sample fractions [11]. When computing risk estimates, ARX must numerically solve equation systems. The solver used by ARX can be configured in the settings dialog. Here, you may specify options such as the total number of iterations, the maximal number of iterations per try and the required accuracy. This may influence the precision of results and execution times. Moreover, in the lower-left corner, the view displays basic risk estimates which are either computed from the sample itself or with a risk model.

Further views that are not shown in the screenshot support analyses of re-identification risks associated with groups of attributes. This analysis uses estimates of uniqueness either based on the sample or on a statistical model. When

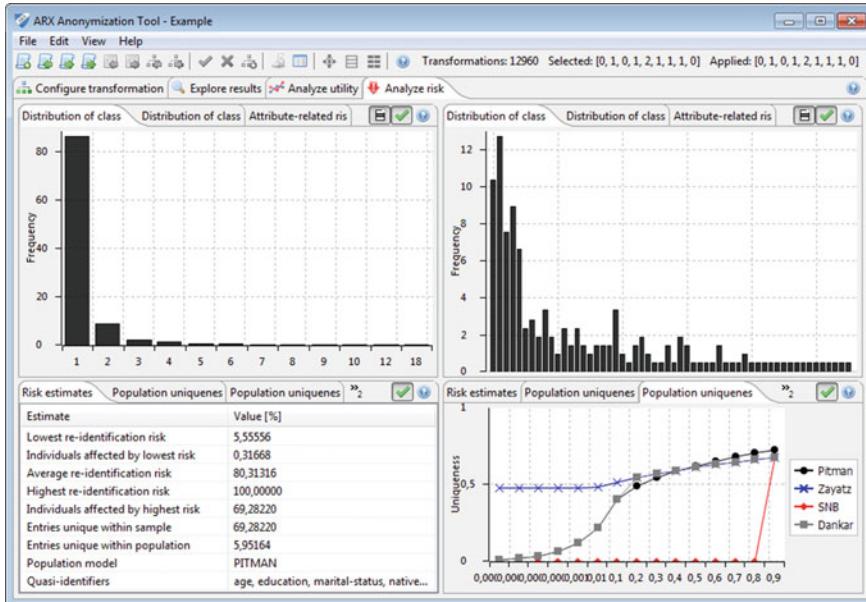


Fig. 6.12 The ARX risk analysis perspective

a set of attributes has been selected for analysis, ARX will determine the average risk of re-identification that is associated with the power set of these attributes. This information can be helpful to decide which attributes need to be generalized. Moreover, details about the population from which the dataset has been sampled can be configured in a dedicated view.

6.3 Implementation Details

The three-step workflow implemented by ARX poses considerable challenges in terms of efficiency. As the process likely involves repeated anonymization with different parameters and configurations, it is important that it is fast. For this purpose, our tool is not built upon existing database systems but implements a dedicated runtime environment that is tailored to the problem domain. Moreover, ARX features a carefully designed search algorithm that employs multiple pruning strategies and it uses efficient implementations of computationally complex methods. In previous work, we have shown that our search algorithm outperforms comparable algorithms within our runtime environment [26, 44].

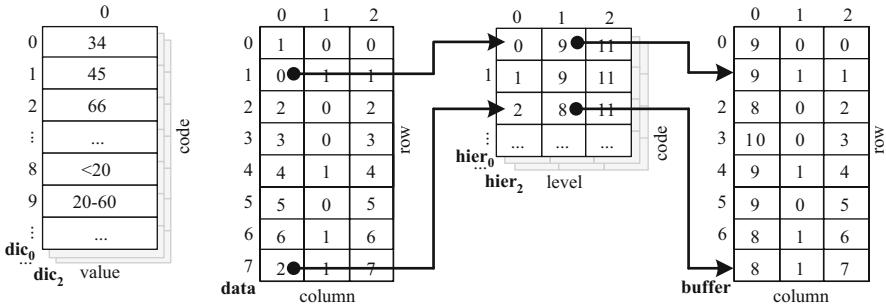


Fig. 6.13 Example of how data is encoded and transformed in ARX

6.3.1 Data Management

The internal encoding and representation of data is one of the key reasons for ARX's scalability. Original and transformed values are encoded into data dictionaries with one dictionary per attribute. As a consequence the internal modules of ARX almost entirely process 32-bit integer values (methods that need to interpret values use the dictionary to access the actual content associated with integer identifiers). For each quasi-identifier, the original values from the input dataset are encoded before the values contained in the higher levels of the generalization hierarchy. As a result, the original attribute values of a quasi-identifier with m distinct values are represented by the numbers 0 to $m - 1$. This allows representing generalization hierarchies as two-dimensional arrays in which the i -th row ($0 \leq i \leq m - 1$) represents the iterative generalization steps for the attribute value encoded with i . This means that the j -th column stores the corresponding transformed value at the j -th level of the hierarchy. The resulting memory layout is sketched in Fig. 6.13.

The input dataset itself is also represented as an integer array (*data*). Additionally, a similar data structure (*buffer*) is maintained which is used to store a transformed representation of the original data. Based on this memory layout, the transformation of a value from the input data in cell (*row*, *col*) to the value defined on level *level* of its generalization hierarchy and storing it in the buffer can be implemented with a simple assignment:

$$\text{buffer}[\text{row}, \text{col}] \leftarrow \text{hier}[\text{col}, \text{data}[\text{row}, \text{col}], \text{level}]$$

The example in Fig. 6.13 shows the transformation of the values in the first column of rows 0 and 7 to the first generalization level. More details about the coding scheme can be found in [26, 27].

The second key to efficient data processing in ARX is its *screenshotting* functionality. Here, the basic idea is to create snapshots that completely represent the state of the dataset for a given transformation while being highly compressed in comparison to the transformed dataset itself. Due to the monotonicity of

generalization hierarchies in ARX, these compressed representations can be used as a basis for further generalization (see Sect. 6.3.2). A related method has been called *roll-up* in previous work [30]. In ARX, a snapshot of a dataset contains a list of its equivalence classes, where each equivalence class is represented by the index of one of its data entries (called *representative*) as well as its size (in terms of the number of entries in the class). Additionally, each equivalence class may also be associated with references to frequency distributions of values of attributes that are not part of the attributes that are generalized. Such attributes are, for example, sensitive attributes for which ℓ -diversity or t -closeness is enforced or quasi-identifiers to which microaggregation is applied.

When a snapshot is used as a basis for generalization, only the representatives need to be transformed and grouped. Additionally, if two or more equivalence classes are merged because of generalization, their associated frequency distributions must be merged as well. Processing compressed snapshots is much more efficient than processing the data itself. ARX employs several strategies for determining whether a snapshot should be created because it is likely to be used in further processing. Moreover, it implements a LRU eviction policy to control memory consumption. Figure 6.14 shows an example of how snapshots are constructed.

The left side of the figure shows the data which is to be snapshotted. We assume that the first three attributes are quasi-identifiers which are generalized, and that the fourth attribute is sensitive and requires frequency distributions to be computed. Moreover, the data represents the transformation $(0, 1, 0)$ of an input dataset which is not shown. The snapshot will also contain this meta-information. In its uncompressed form, the data is represented by a two-dimensional array which consumes

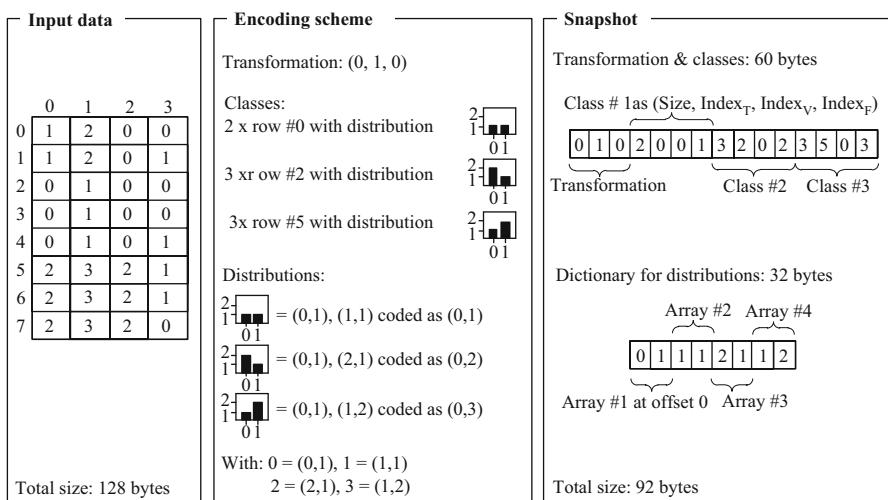


Fig. 6.14 Example of how data snapshots are represented in ARX

$8 * 4 * 4 = 128$ bytes of memory. The data consists of three equivalence classes for which representatives and sizes are contained in the snapshot. Additionally, each class is associated with a frequency distribution. Each frequency distribution consists of two arrays, one that represents the values (i.e., their integer codes) and one that represents the associated counts. These arrays are also encoded into a dictionary, meaning that each array can be referenced by an integer identifier and that arrays that appear multiple times only need to be stored once.

The snapshot consists of the transformation followed by four identifiers per equivalence class: (1) its size, (2) the index of the representative, (3) the identifier of the value array of its distribution, and (4) the identifier of the counts array of its distribution. The resulting representation requires $15 * 4 + 8 * 4 = 92$ bytes of memory. We note that the size of a snapshot is independent of the number of quasi-identifiers and that compression will be more effective in real-world scenarios.

6.3.2 Pruning Strategies

The general process that is implemented by ARX and its search algorithm, *Flash* [26], as well as by other globally-optimal anonymization algorithms that use a similar coding model, such as Incognito [29] and OLA [16], is very simple. To determine an optimal solution, each transformation in the solution space is checked for whether it is a solution candidate, meaning that it fulfills the given set of privacy criteria. Additionally, data utility is computed for all solution candidates, which allows to determine the global optimum. For an individual transformation, this is implemented in the following manner in ARX:

- **Step 1:** Apply the transformation to the dataset.
- **Step 2:** Suppress all entries in all equivalence classes that do not fulfill the privacy model.
- **Step 3:** If the number of suppressed entries is lower than the given threshold, the transformation is a solution candidate.
- **Step 4:** If the transformation is a solution candidate, compute the utility of the transformed dataset. This step may involve microaggregation.

As this is a computationally complex process, anonymization algorithms implement pruning strategies to exclude parts of the search space. In this context, *monotonicity* is a typical assumption. This means that generalizations of a transformation that fulfills a privacy model also fulfill the privacy model. Moreover, specializations of a transformation that does not fulfill a privacy model do also not fulfill the privacy model. For utility measures, monotonicity means that information loss increases monotonically with generalization and therefore the utility of a transformation will always be lower than, or equal to, the utility of its specializations. Monotonicity (of both privacy criteria and utility metrics) requires generalization hierarchies to be monotonic, which means that they must form mono-hierarchies, where each element has exactly one parent. This is guaranteed by the ARX system.

However, in real-world settings monotonicity is only given for simple privacy models, e.g., k -anonymity and distinct- ℓ -diversity, or when the coding model is restricted to global recoding with full-domain generalization. More elaborate privacy model that enforce restrictions on the distributions of sensitive attribute values in each class, e.g., recursive- (c, ℓ) -diversity or t -closeness, are not monotonic with tuple suppression. When generalizing a representation of a dataset previously suppressed data entries can re-join to the dataset and increase the distance of a distribution of sensitive attribute values in a class from the overall distribution (t -closeness) or make a distribution less diverse. The same is true for more elaborate utility measures. An increase in generalization may lead to a decrease in suppression, which increases overall data utility.

The key to ARX's scalability and efficiency when dealing with complex privacy problems is the exploitation of multiple pruning strategies. Firstly, ARX is able to prune parts of the solution space that can never be a solution to the privacy problem, if the set of specified criteria induces a *minimal equivalence class size*. This is the case if the set of privacy criteria contains k -anonymity or any variant of ℓ -diversity [17]. The property of a transformation not resulting in the required minimal class size is equal to the property of not fulfilling k -anonymity, which is monotonic even with tuple suppression [50]. As a consequence, specializations of transformations that do not have this property can be pruned from the search process.

Secondly, ARX employs a generic pruning strategy that is based on data utility measures and which is therefore independent of the privacy model utilized. Here, ARX computes two different variants of the selected utility measure for each transformation: (1) a utility measure only considering generalization (termed *upper bound*) and (2) a utility measure that considers generalization and suppression. It is easy to see that utility measured when only generalizing a dataset is always greater than or equal to the utility measured when first generalizing a dataset and then suppressing some of its entries. Moreover, all utility measures implemented in ARX are monotonic for coding models that only consist of generalization (see [3, 16, 24, 30]). During the search process, ARX keeps track of the current optimum, i.e., the transformation with maximal utility. Transformations can be pruned if the upper bound of the utility of any of its specializations, i.e., predecessors in the lattice, is already lower than or equal to the current optimum [3]. As we will show in Sect. 6.4, this method is a highly effective.

To implement these pruning strategies into the search process, the ARX tool employs six different variants of the *Flash* algorithm [26] that utilize different combinations of the above and previously proposed pruning strategies. An appropriate variant of the algorithm is chosen depending on the exact configuration of a given anonymization problem.

6.3.3 Risk Analysis and Risk-Based Anonymization

ARX supports arbitrary combinations of privacy criteria, including risk-based criteria that ensure that estimated re-identification risks fall below a given threshold. For these criteria, ARX uses a coding model that integrates well with the other functionalities of the system and for which it is possible to determine an optimal solution to the anonymization problem. To transform a dataset according to a set of privacy criteria that contains one or more risk-based criteria the process described in Sect. 6.3.2 is modified slightly:

- **Step 1:** Apply the transformation to the dataset.
- **Step 2:** Suppress all entries in all equivalence classes that do not fulfill the (not risk-based) privacy criteria.
- **Step 3:** While risk estimates are greater than the given thresholds:
 - **Step 3.1:** Suppress all entries in the equivalence class with the lowest information content (determined with the selected utility measure).
 - **Step 3.2:** Re-evaluate the risk model.
- **Step 4:** If the number of suppressed entries is lower than the given threshold, the transformation is a solution candidate.
- **Step 5:** If the transformation is a solution candidate, compute the utility of the transformed dataset.

ARX implements three different statistical models to estimate population uniqueness, two of which require solving *bi-variate non-linear equation systems*. In ARX, we need efficient implementations of this process, as risk models will be evaluated frequently during the above search process. For solving the equation systems, ARX implements Newton's method [56], which requires two *object functions* for which a common root is to be found as well as four partial derivatives, two for each object function. In both super-population models the object function is a combination of sums over the sizes of equivalence classes in the dataset and therefore computationally complex to evaluate. To overcome this bottleneck we have implemented the object functions as well as the partial derivatives with the family of *polygamma* functions [57], resulting in closed forms of the open sum formulas.

Moreover, there are efficient numeric approximations of polygamma functions. In ARX, we use an implementation developed at Microsoft Research [38]. As approximations of polygamma are inaccurate for some ranges of input, we employ a three step-process:

- **Step 1:** Try to solve the linear equation system with the approximations of all input functions.
- **Step 2:** If a result has been found, verify it with the original open forms of the object functions.
- **Step 3:** If the verification fails or no solution has been found by step 1, try to solve the system using the original open forms of the object functions.

We confirmed experimentally that this scheme is orders of magnitude faster than solving the equation systems with the open forms of the object functions. The final verification step ensures that no errors are introduced by using numeric approximations. The described scheme results in efficient search processes for risk-based anonymization and enables risk analyses that can be performed in near real-time.

6.4 Experimental Evaluation

In this section we will briefly evaluate our system with a specific focus on the methods and implementation details presented in the previous sections. To this end, we make use of the benchmarking environment that we have presented in [44]. The implementation of the experiments performed in this section is also available online [43]. To cover a broad spectrum of privacy methods supported by ARX we present two evaluations. Firstly, we anonymize datasets with k-anonymity, which is the least complex privacy criterion available in our tool. Secondly, we anonymize datasets with a risk-based criterion that makes use of the decision rule by Dankar et al. As this rule depends on three different statistical models and risk-based criteria utilize a more complex coding model, this is the most complex privacy model supported by ARX. Moreover, risk-based anonymization is also a relaxed variant of k-anonymity, making a comparison between the data utility resulting from these two models interesting. As a utility measure we used *Loss*, which is also the most elaborate method that is available for this purpose in ARX. Additionally, we make use of the tool's feature of automatically balancing generalization and suppression to achieve optimal data utility by setting the suppression limit to 100 %.

Our benchmark is based upon five real-world datasets, which we have chosen for different reasons. Two of the datasets are frequently used for benchmarking work on data anonymization and thus represent de-facto standard problem instances: the 1994 US census database (Adult) and the dataset for the 1998 KDD data mining competition (Cup). We include three additional datasets with increasing size: US NHTSA crash statistics (FARS), American Time Use Survey (ATUS) and Integrated Health Interview Series (IHIS). The datasets feature different numbers of data entries (ranging from 30,162 records (Adult) to 1,193,504 records (IHIS)). Moreover, they consist of eight and nine quasi-identifiers. As can be seen in Table 6.2, the search spaces consisted of between 12,960 and 45,000 transformations. More details about the datasets and the generalization hierarchies utilized can be found in [44].

Table 6.3 shows runtime measures obtained for risk-based anonymization with a threshold of 1 % population uniqueness according to Dankar et al.'s model. In this case, the only pruning strategy that can be utilized is to exclude transformations based on upper bounds of data utility. It can be seen that this strategy is highly effective, as more than 99 % of the search space were excluded for all datasets. The execution times, however, were rather high with totals between 2 and 216 s.

Table 6.2 Datasets used in the experiments

Dataset	QIs	Entries	Transformations	Size [MB]
Adult	9	30,162	12,960	2.52
Cup	8	63,441	45,000	7.11
FARS	8	100,937	20,736	7.19
ATUS	9	539,253	34,992	84.03
IHIS	9	1,193,504	25,920	107.56

Table 6.3 Runtime measures for risk-based anonymization

Dataset	Trans. checked	Trans. pruned [%]	Time total [s]	Time per check [ms]
Adult	61	99.53	2	34
Cup	431	99.04	55	128
FARS	165	99.20	30	183
ATUS	19	99.95	21	1,092
IHIS	97	99.63	216	2,229

Table 6.4 Runtime measures for 5-anonymity

Dataset	Trans. checked	Trans. pruned [%]	Time total [s]	Time per check [ms]
Adult	375	97.11	1	3
Cup	1,245	97.23	13	11
FARS	1,102	94.69	8	7
ATUS	130	99.63	4	34
IHIS	507	98.04	65	128

This is due to the complexity of the coding model for risk-based privacy criteria. This complexity is also reflected by the average time required for each individual check, which was between 34 ms (Adult) and 2.2 s (IHIS).

Table 6.4 shows runtime measures obtained for anonymization the datasets with 5-anonymity. In contrast to the previous experiment, multiple pruning strategies can be utilized as k-anonymity is a monotonic model. Still, pruning was less effective being as low as 94 % for the FARS dataset. The execution times, however, were much lower with totals between 1 and 65 s. The reason is that it is less complex to check data transformations for k-anonymity. Again, this is also reflected by the average time required for each individual check, which was between 3 (Adult) and 128 ms (IHIS).

The privacy models used in this section result in very different privacy guarantees. 5-anonymity ensures that the re-identification risk for each entry from the dataset is lower than or equal to 20 %. The risk used as a basis for the anonymization is likely to be strongly overestimated. The risk-based criterion ensures that at most 1 % of the entries from the dataset can be re-identified. This is a much less strict criterion that may, however, still offer sufficient protection, e.g. when assuming a *prosecutor scenario* where the attacker aims at re-identifying the data of one specific individual. The difference in protection is also reflected by the amount

Table 6.5 Utility measures for risk-based anonymization

Dataset	Entries suppressed	Attributes generalized	Generalization levels	Information loss [%]
Adult	0 (0 %)	2/9 (22 %)	1/4, 1/3	2.40
Cup	0 (0 %)	3/8 (38 %)	3/5, 1/4, 1/4	2.68
FARS	0 (0 %)	3/8 (38 %)	3/5, 1/3, 1/3	6.55
ATUS	4,160 (1 %)	1/9 (11 %)	1/5	1.00
IHIS	0 (0 %)	3/9 (33 %)	1/5, 1/3, 1/4	1.30

Table 6.6 Utility measures for 5-anonymity

Dataset	Entries suppressed	Attributes generalized	Generalization levels	Information loss [%]
Adult	3,382 (11 %)	5/9 (56 %)	3/4, 1/3, 1/2, 1/2	22.85
Cup	2,578 (4 %)	5/8 (63 %)	4/5, 3/4, 1/4, 2/4, 1/4	13.97
FARS	4,636 (5 %)	6/8 (75 %)	3/5, 1/2, 1/3, 1/3, 1/2	17.58
ATUS	64,007 (12 %)	2/9 (22 %)	2/5, 1/2	12.71
IHIS	182,663 (15 %)	3/9 (33 %)	2/5, 1/3, 2/4	17.28

of generalization and suppression that needs to be applied to the dataset. We will analyze these differences in the following paragraphs.

Table 6.5 gives an overview of how the data needed to be transformed to achieve optimality for the risk-based model. It can be seen that suppression was only required for one dataset (ATUS). Moreover, only between one and three attributes needed to be generalized for each dataset. If an attribute was generalized, it was only transformed to the lowest generalization level in 10 out of 12 cases. The rightmost column shows the information loss according to the normalized Loss utility measure. It can be seen that the loss of information was evaluated to be between only 1 and 7 % by this utility measure.

Table 6.6 shows the resulting utility for 5-anonymity. It can be seen that this model is much more difficult to achieve. Suppression was required for all datasets, while between 4 and 15 % of the entries needed to be removed for the optimal solution. Additionally, on average, half of the attributes in a dataset needed to be generalized. Still, if an attribute was generalized, it was transformed to a rather low generalization level in most cases. The loss of information was evaluated to be between 13 and 23 % for this configuration.

The numbers reflecting data utility also explain why the pruning power was lower in the second experiment. As the datasets obviously required much less generalization with risk-based anonymization, the search strategy was able to utilize a much stronger upper-bound for pruning transformations based on data utility. In the second experiment, more generalization was required, which reduced pruning power. This was compensated to some extent by pruning methods that use the monotonicity of k-anonymity.

6.5 Discussion

In this section, we will compare ARX to other data anonymization tools, discuss its limitations and present directions for future work. We will conclude this article with some general remarks about data anonymization.

6.5.1 Comparison with Prior Work

The landscape of existing tools is heterogeneous. To our knowledge, PARAT is the only commercial de-identification software using syntactic privacy models [46]. It is a closed-source tool for which only limited information is available to the public. It uses the same coding model as ARX and a related search strategy [16]. We will focus on non-commercial tools in the remainder of this article.

Several research prototypes exist that have mainly been developed for demonstration purposes. Problems with these tools include scalability issues when handling large datasets, complex configuration requiring IT expertise, and incomplete support of methods for statistical disclosure control. Our tool is the only software that provides wizards for creating generalization hierarchies and visualizations of the solution space.

The UTD Anonymization Toolbox (UTD-AT) is written in Java. It supports three different privacy models (k -anonymity, ℓ -diversity and t -closeness) [54]. When using the k -anonymity privacy model it supports the same coding model as ARX. The tool implements Incognito, which is a globally-optimal search algorithm [29], and DataFly, which is a heuristic search algorithm [49]. In [26, 45] we have shown that ARX’s search algorithm significantly outperforms Incognito within our runtime environment. When using ℓ -diversity or t -closeness, data transformation is restricted to full-domain generalization, because the tool is not able to handle non-monotonic privacy problems. This can lead to low data quality [3]. UTD-AT supports two further coding models: multi-dimensional global recoding with generalization via the Mondrian algorithm [31] and anatomization [58], which is a specific form of slicing [34]. The tool uses a SQLite database backend and we encountered scalability issues with larger datasets. It does not provide a graphical interface and requires configuration to be performed via an XML file. As the name already implies, the software only provides a loosely coupled collection of methods that are not harmonized and integrated in a comprehensive manner. The tool does not implement methods for risk analyses or risk-based anonymization and it only supports few methods for measuring data utility.

The Cornell Anonymization Tool (CAT) is implemented in C++ [59]. It implements ℓ -diversity and t -closeness. As a coding model it uses global recoding with full-domain generalization. It only supports manual tuple suppression. For importing data it requires files to be in a tool-specific format. Analogously to UTD-AT it implements the Incognito algorithm [29]. It only provides few methods for automatically evaluating data utility and for assessing disclosure risks.

TIAMAT is a closed source software that is implemented in Java [9]. It only supports the k -anonymity privacy model. It implements the Mondrian [31] partitioning algorithm and k -Member [4], which is a clustering algorithm. The coding model is multi-dimensional global recoding with generalization. The tool implements a simple graphical editor for generalization hierarchies. For automatically measuring data utility, TIAMAT implements the Global Certainty Penalty (GCP) method [19] and the Classification Metric [24]. The aim of the tool is to compare the Mondrian algorithm with the k -Member approach and it therefore implements various visualizations that summarize their execution times and resulting data utility. The tool does not implement methods for risk assessment.

SECRETA is a closed source software that is implemented in C++ [42]. It has been developed for comparing methods for anonymizing relational and transactional (set-valued) data. For relational data it implements a clustering algorithm, top-down and bottom-up search algorithms and Incognito [29]. Coding models for relational data include full-domain generalization, subtree generalization and local recoding [17]. It further implements five different algorithms for anonymizing set-valued data, including the method from [41], and three different methods for combining algorithms for relational data with algorithms for set-valued data. Analogously to ARX and CAT, the user interface provides visualizations of data characteristics with methods from empirical statistics. The tool implements a sophisticated method for measuring data utility, by specifying query workloads and determining the average relative error of query results [60]. As the aim of the tool is to compare different anonymization strategies, it further implements various visualizations of execution times and data utility. It does not provide methods for risk assessment and it does not enable users to visualize and browse a solution space.

In addition to these research prototypes, there are two mature solutions that are being used in practice. Both have been developed by the statistics research community: sdcMicro and μ -Argus. Compared to ARX, both implement a more manual anonymization process. In terms of supported methods, they provide more features in some areas of statistical disclosure control (e.g. more coding models) and less features in others (e.g. fewer privacy models).

sdcMicro is an open source package for the R statistics software parts of which have been implemented in C++ [51]. It supports many primitives required for data anonymization. The tool features a graphical user interface, which only provides access to a subset of the methods implemented by the package. sdcMicro is not designed as a stand-alone application, but it is oriented towards statistics experts that use the console to leverage the full potential of the package. As the tool is integrated into R, a wide variety of methods for analyzing and visualizing data can be used. Although the most important usage scenario of the tool is a manual iterative process of data anonymization, it also supports methods for automatically transforming data with local recoding or top-coding. It supports two syntactic privacy models: k -anonymity and ℓ -diversity. sdcMicro implements a wide variety of truthful and perturbative transformation models, including noise addition, global recoding, local suppression, microaggregation, post-randomization, rank-swapping as well as top-and bottom-coding [8, 17]. Although it also supports the same coding model as

ARX, the tool is focused on a manual process and it does not implement algorithms for automatically constructing and analyzing a solution space. Consequently, it only implements few utility measures. It supports various risk estimators, some of which are based on super-population models.

μ -Argus is a former closed-source application, which has just recently been converted to an open source project after development had ceased for several years [23]. Its core functionalities are implemented in C++ and, recently, a graphical user interface has been developed in Java. Similar to sdcMicro it offers only few methods that assist users in the data anonymization process and it does not support automated recoding. It provides a broad spectrum of recoding techniques, including global recoding with local suppression as well as top- and bottom-coding and multiple methods for risk estimation. De-identification must be performed manually in a three-step process. First a set of attributes is selected as a quasi-identifier. It is recommended that these sets do not contain more than about a handful of attributes. Next, re-identification risks are analyzed. Finally, the user performs recoding operations until risks fall below an acceptable threshold. This process can then be repeated with another set of attributes. The tool does not implement typical utility measures and provides only few visualizations.

The syntactic privacy models implemented by ARX and related software can be an important building block for protecting sensitive health data. However, they can only provide very specific guarantees which require rather strong assumptions to be made about the goals and the background knowledge of an attacker. Semantic privacy models, such as differential privacy [13], require much less such assumptions. In contrast to syntactic models, which formulate syntactic conditions for released datasets, semantic models are directly related to a formalization of privacy [12]. However, only few implementations are available, e.g., PINQ [37] or HIDE [18], most of which focus on privacy-preserving data analysis and thus on a usage scenario that is different from the one considered by our work. Additionally, these methods involve stronger trade-offs in terms of supported workflows [10]. Many methods of differential privacy are non-truthful, which can be problematic for an application in the biomedical domain [10]. Which approach provides a better balance between privacy and utility is subject to ongoing discussions (see, e.g. [5, 39]).

6.5.2 Limitations and Future Work

ARX currently relies on a globally-optimal search strategy and materializes the complete solution space. As a consequence, it can only handle anonymization problems in which the search space is small enough to allow materialization. As a rule of thumb, the current version of ARX can handle datasets with up to about 15 quasi-identifiers, but the exact limitation depends on the size of the generalization hierarchies utilized. In the near future, we will integrate a heuristic search strategy for anonymizing high-dimensional datasets.

While ARX already covers a broad spectrum of methods, it does currently not support methods for handling set-valued data [17]. In future work, we will implement methods for anonymizing data with set-valued attributes, such as k^m -anonymity [52], and methods for anonymizing data with set-valued attributes and relational characteristics, such as (k, k'') -anonymity [41]. We further plan to adapt the method of measuring data utility with predefined query workloads to our setup [60].

We are also working on the integration of semantic privacy models into the tool. Here, we focus on non-interactive methods of Differential Privacy (DP) [13]. For example, (k, β) -SDGS is very well suited for integration into ARX [33]. It implements a combination of random sampling, k -anonymization with generalization and suppression to create datasets that fulfill (ϵ, δ) -DP [14]. These primitives for data transformation are already implemented in our tool.

6.5.3 Concluding Remarks

While ARX aims at making data anonymization available to a wide variety of users, data anonymization remains a complex issue that has to be performed by experts. There is no single measure that is able to protect datasets from all possible threats, especially while being flexible enough to support many usage scenarios. As is common in IT security, data controllers should therefore follow the *onion layer principle* and employ a multitude of measures for protecting sensitive personal data. This includes legal agreements, as well as *data economy*, meaning the principle that no more personal details than necessary should be collected, stored and shared.

Acknowledgements The authors would like to express their appreciation to Klaus A. Kuhn for his many helpful and insightful comments and suggestions.

References

1. Article 29 Data Protection Working Party: Opinion 05/2014 on anonymisation techniques. http://www.cnpd.public.lu/fr/publications/groupe-art29/wp216_en.pdf. Accessed 22 Apr (2014)
2. ARX – Powerful Data Anonymization: <http://arx.deidentifier.org/>. Accessed 06 May (2015)
3. Bayardo, R., Agrawal, R.: Data privacy through optimal k -anonymization. In: Proceedings of the International Conference on Data Engineering, pp. 217–228 (2005)
4. Byun, J., Sohn, Y., Bertino, E., Li, N.: Secure anonymization for incremental datasets. In: Proceedings of VLDB Workshop Secure Data Management, pp. 48–63 (2006)
5. Cavoukian, A., Castro, D.: Big data and innovation, setting the record straight: de-identification does work. Privacy by Design, Ontario, Canada. <http://www2.itif.org/2014-big-data-deidentification.pdf> (2014). Accessed 06 May (2015)
6. Chen, G., Keller-McNulty, S.: Estimation of identification disclosure risk in microdata. *J. Off. Stat.* **14**, 79–95 (1998)

7. Ciglic, M., Eder, J., Koncilia, C.: k-anonymity of microdata with null values. In: Proceedings of International Conference on Database and Expert Systems Applications (2014)
8. Ciriani, V., De Capitani di Vimercati, S., Foresti, S., Samarati, P.: Microdata protection. In: Yu, T., Jajodia, S. (eds.) *Secure Data Management in Decentralized Systems. Advances in Information Security*, vol. 33, pp. 291–321. Springer, Berlin (2007)
9. Dai, C., Ghinita, G., Bertino, E., Byun, J.W., Li, N.: TIAMAT: a tool for interactive analysis of microdata anonymization techniques. In: Proceedings of the VLDB Endowment (2009)
10. Dankar, F.K., Emam, K.E.: Practicing differential privacy in health care: a review. *Trans. Data Privacy* **6**(1), 35–67 (2013)
11. Dankar, F., Emam, K.E., Neisa, A., Roffey, T.: Estimating the re-identification risk of clinical data sets. *BMC Med. Inform. Decis. Mak.* **12**(1), 66 (2012)
12. Dwork, C.: An ad omnia approach to defining and achieving private data analysis. In: Proceedings of PinKDD, pp. 1–13 (2007)
13. Dwork, C.: Differential privacy. In: *Encyclopedia of Cryptography and Security*, pp. 338–340. Springer, Berlin (2011)
14. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: privacy via distributed noise generation. In: Proceedings of EUROCRYPT 2006, pp. 486–503 (2006)
15. El Emam, K., Jonker, E., Arbuckle, L., Malin, B.: A systematic review of re-identification attacks on health data. *PloS One* **6**(12), e28071 (2011)
16. Emam, K.E., Dankar, F.K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J., Walker, M., Chowdhury, S., Vaillancourt, R., Roffey, T., Bottomley, J.: A globally optimal k-anonymity method for the de-identification of health data. *J. Am. Med. Inform. Assoc.* **16**(5), 670–682 (2009)
17. Fung, B., Wang, K., Fu, A., Yu, P.: *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. CRC Press, Hoboken (2010)
18. Gardner, J.J., Xiong, L., Li, K., Lu, J.J.: HIDE: heterogeneous information de-identification. In: Proceedings of International Conference on Extending Database Technology, pp. 1116–1119 (2009)
19. Ghinita, G., Karras, P., Kalnis, P., Mamoulis, N.: Fast data anonymization with low information loss. In: Proceedings of the VLDB Endowment, pp. 758–769 (2007)
20. Gkoulalas-Divanis, A., Loukides, G., Sun, J.: Publishing data from electronic health records while preserving privacy: a survey of algorithms. *J. Biomed. Inform.* **50**, 4–19 (2014)
21. Greenberg, B., Zayatz, L.: Strategies for measuring risk in public use micro-data files. *Statistica Neerlandica* **46**(1), 33–48 (1992)
22. Hoshino, N.: Applying Pitman's sampling formula to microdata disclosure risk assessment. *J. Off. Stat.* **17**(4), 499–520 (2001)
23. Hundepool, A., van de Wetering, A., Ramaswamy, R., Franconi, L., Polettini, S., Capobianchi, A., de Wolf, P.P., Domingo, J., Torra, V., Brand, R., Giessing, S.: μ -Argus manual. <http://neon.vb.cbs.nl/casc/Software/MuManual4.2.pdf>. Accessed 22 Apr (2008)
24. Iyengar, V.: Transforming data to satisfy privacy constraints. In: Proceedings of International Conference on Knowledge Discovery and Data Mining, pp. 279–288 (2002)
25. Kayaalp, M., Browne, A.C., Dodd, Z., Sagan, P., McDonald, C.: De-identification of address, date, and alphanumeric identifiers in narrative clinical reports. In: AMIA Annual Symposium Proceedings, pp. 767–776 (2014)
26. Kohlmayer, F., Prasser, F., Eckert, C., Kemper, A., Kuhn, K.A.: Flash: efficient, stable and optimal k-anonymity. In: Proceedings of International Conference on Information Privacy, Security, Risk and Trust (2012)
27. Kohlmayer, F., Prasser, F., Eckert, C., Kemper, A., Kuhn, K.A.: Highly efficient optimal k-anonymity for biomedical datasets. In: Proceedings of International Symposium on Computer-Based Medical Systems (2012)
28. Kohlmayer, F., Prasser, F., Eckert, C., Kuhn, K.A.: A flexible approach to distributed data anonymization. *J. Biomed. Inform.* **50**, 62–76 (2013).
29. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: efficient full-domain k-anonymity. In: Proceedings of International Conference on Management of Data, pp. 49–60 (2005)

30. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Multidimensional k-anonymity (TR-1521). Tech. Rep., University of Wisconsin (2005)
31. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: Proceedings of International Conference on Data Engineering, p. 25 (2006)
32. Li, N., Li, T., Venkatasubramanian, S.: t -closeness: privacy beyond k -anonymity and ℓ -diversity. In: Proceedings of International Conference on Data Engineering, pp. 106–115 (2007)
33. Li, N., Qardaji, W.H., Su, D.: Provably private data anonymization: or, k -anonymity meets differential privacy. CoRR, abs/1101.2604 **49**, 55 (2011)
34. Li, T., Li, N., Zhang, J., Molloy, I.: Slicing: a new approach for privacy preserving data publishing. Trans. Knowl. Data Eng. **24**(3), 561–574 (2012)
35. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: ℓ -diversity: privacy beyond k -anonymity. Trans. Knowl. Discov. Data **1**(1), 24–35 (2007)
36. Malin, B., Benitez, K., Masys, D.: Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA privacy rule. J. Am. Med. Inform. Assoc. **18**(1), 3–10 (2011)
37. McSherry, F.: Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proceedings of International Conference on Management of Data, pp. 19–30 (2009)
38. Minka, T.: Lightspeed Matlab toolbox. <http://research.microsoft.com/en-us/um/people/minka/software/lightspeed/>. Accessed 22 Apr (2014)
39. Narayanan, A., Felten, E.: No silver bullet: de-identification still doesn't work. <http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf> (2014). Accessed 06 May (2015)
40. Nergiz, M.E., Atzori, M., Clifton, C.: Hiding the presence of individuals from shared databases. In: Proceedings of International Conference on Management of Data, pp. 665–676 (2007)
41. Poulis, G., Loukides, G., Gkoulalas-Divanis, A., Skiadopoulos, S.: Anonymizing data with relational and transaction attributes. In: Proceedings of ECML PKDD, pp. 353–369 (2013)
42. Poulis, G., Gkoulalas-Divanis, A., Loukides, G., Skiadopoulos, S., Tryfonopoulos, C.: SEC-RETA: a system for evaluating and comparing relational and transaction anonymization algorithms. In: Proceedings of International Conference on Extending Database Technology, pp. 620–623 (2014)
43. Prasser, F., Kohlmayer, F.: A simple benchmark of risk-based anonymization with ARX. <https://www.github.com/arkx-deidentifier/risk-benchmark>. Accessed 22 Apr (2015)
44. Prasser, F., Kohlmayer, F., Kuhn, K.A.: A benchmark of globally-optimal anonymization methods for biomedical data. In: Proceedings of International Symposium on Computer-Based Medical Systems (2014).
45. Prasser, F., Kohlmayer, F., Lautenschlaeger, R., Eckert, C., Kuhn, K.A.: ARX: a comprehensive tool for anonymizing biomedical data. In: AMIA Annual Symposium Proceedings (2014).
46. Privacy Analytics Inc.: About PARAT de-identification software. <http://www.privacyanalytics.ca/software/parat/>. Accessed 22 Apr (2015)
47. Rinott, Y.: On models for statistical disclosure risk estimation. In: Proceedings of ECE/Eurostat Work Session on Statistical Data Confidentiality, pp. 275–285 (2003)
48. Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information. In: Proceedings of Symposium on Principles of Database Systems, p. 188 (1998)
49. Sweeney, L.: Datafly: a system for providing anonymity in medical data. In: Database Security, XI: Status and Prospects, p. 20 (1998)
50. Sweeney, L.: Computational disclosure control: a primer on data privacy protection. Ph.D. thesis, MIT (2001)
51. Templ, M.: Statistical disclosure control for microdata using the r-package sdcmicro. Trans. Data Privacy **1**(2), 67–85 (2008)
52. Terrovitis, M., Mamoulis, N., Kalnis, P.: Privacy-preserving anonymization of set-valued data. In: Proceedings of the VLDB Endowment (2008)

53. U.S. Health Insurance Portability and Accountability Act of 1996. Public Law 1-349 (1996)
54. UTD Data Security and Privacy Lab: UTD anonymization toolbox. <http://www.cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php>. Accessed 10 June (2012)
55. Wikipedia: Hasse diagram. https://en.wikipedia.org/wiki/Hasse_diagram. Accessed 22 Apr (2015)
56. Wikipedia: Newton's method. https://en.wikipedia.org/wiki/Newton's_method. Accessed 22 Apr (2015)
57. Wikipedia: Polygamma function. https://en.wikipedia.org/wiki/Polygamma_function. Accessed 22 Apr (2015)
58. Xiao, X., Tao, Y.: Anatomy: simple and effective privacy preservation. In: Proceedings of the VLDB Endowment, pp. 139–150 (2006)
59. Xiao, X., Wang, G., Gehrke, J.: Interactive anonymization of sensitive data. In: Proceedings of International Conference on Management of Data, pp. 1051–1054 (2009)
60. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.W.: Utility-based anonymization using local recoding. In: Proceedings of International Conference on Knowledge Discovery and Data Mining, pp. 785–790 (2006)
61. Zayatz, L.V.: Estimation of the percent of unique population elements on a microdata file using the sample. Statistical Research Division Report Number: Census/SDR/RR-91/08 (1991)

Chapter 7

Utility-Constrained Electronic Health Record Data Publishing Through Generalization and Disassociation

Grigoris Loukides, John Liagouris, Aris Gkoulalas-Divanis, and Manolis Terrovitis

Abstract Data containing diagnosis codes are often derived from electronic health records and shared to enable large-scale, low-cost medical studies. However, the sharing of such data may lead to the disclosure of patients' identities, which must be prevented to address privacy concerns and comply with worldwide legislation. To ensure that data privacy and utility are preserved, a *utility-constrained* anonymization approach can be enforced. This approach transforms a given dataset, so that the probability of identity disclosure, based on diagnosis codes, is limited and the data remain useful for intended studies. In this chapter, we provide a detailed discussion of the utility-constrained anonymization approach. Specifically, we explain how utility constraints, which model the requirements of intended studies, can be formulated and satisfied through data generalization or disassociation. Furthermore, we review two recently proposed algorithms that follow the utility-constrained approach and are the current state-of-the-art in terms of preserving data utility. We conclude this chapter by discussing several promising directions for future research.

G. Loukides (✉)

School of Computer Science and Informatics, Cardiff University, Cardiff, UK

e-mail: g.loukides@cs.cf.ac.uk

J. Liagouris

Department of Electrical and Computer Engineering, National Technical University of Athens, Kaisariani, Greece

e-mail: liagos@dblab.ece.ntua.gr

A. Gkoulalas-Divanis

Smarter Cities Technology Center, IBM Research, Ballsbridge, Ireland

e-mail: arisdiva@ie.ibm.com

M. Terrovitis

Institute for the Management of Information Systems, Research Center Athena, Marousi, Greece

e-mail: mter@imis.athena-innovation.gr

7.1 Introduction

Electronic Health Records (EHRs) allow healthcare professionals to make informed decisions that have a very positive impact on quality of care and patient safety. In addition, EHRs contain large amounts of patient data that can be shared to improve medical science, greatly reduce research costs, and enable large-scale, complex medical studies. For instance, various analytic tasks, ranging from building predictive data mining models [7] to genomic studies [1], can be performed using data derived from EHRs. Thus, there has recently been a tremendous interest for sharing large amounts of electronic health record data [19, 24].

The sharing of EHR data must be performed in a privacy-preserving way, to address patients' concerns and comply with legislation. The requirement to preserve privacy is posed by a number of regulations, such as the HIPAA privacy rule¹ in the United States, the Anonymization Code² in the United Kingdom, and the Data Protection Directive³ in the European Union. However, despite the existence of such legislation, privacy breaches that affect many individuals still occur frequently. Specifically, 627 privacy breaches, which affected more than 500 and up to 4.9 M individuals each, were reported from 2010 to July 2013 by the U.S. Department of Health & Human Services.⁴

7.1.1 Identity Disclosure

An alarming threat related to the sharing of EHR data is *identity disclosure* (also referred to as *re-identification*). An identity disclosure attack succeeds when an identified patient is associated with their record in the shared data. Identity disclosure may occur even when the data are devoid of explicit identifiers (i.e., attributes that directly identify patients, such as phones and social security numbers). This is because explicit identifiers are often publicly available and can be linked to the shared data, based on diagnosis codes.

Consider, for example, the data in Table 7.1a. Each record in the data corresponds to a distinct patient and contains the diagnosis codes of the patient. The ID column is for referencing and is not released, and the description of the diagnosis codes in Table 7.1a is shown in Table 7.1b. Observe that an attacker, who knows that a patient is diagnosed with *Bipolar I disorder, single manic episode, mild* (denoted with the ICD code 296.01) and *Closed dislocation of finger, unspecified part* (denoted with the ICD code 834.0), can associate an identified patient with the first record, denoted

¹<http://www.ncvhs.hhs.gov/091210p06b.pdf>.

²http://ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation.

³<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML>.

⁴<http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrule/breachtool.html>.

Table 7.1 (a) Dataset comprised of diagnosis codes, and (b) diagnosis codes contained in the dataset of Table 7.1a and their description (reprinted from [15], with permission from Elsevier)

ID	Records
r_1	{296.00, 296.01, 296.02, 834.0, 944.01}
r_2	{296.00, 296.02, 296.01, 401.0, 944.01, 692.71, 695.10}
r_3	{296.00, 296.02, 692.71, 834.0, 695.10}
r_4	{296.00, 296.01, 692.71, 401.0}
r_5	{296.00, 296.01, 296.02, 692.71, 695.10}
r_6	{296.03, 295.04, 404.00, 480.1}
r_7	{294.10, 296.03, 834.0, 944.01}
r_8	{294.10, 295.04, 296.03, 480.1}
r_9	{294.10, 295.04, 404.00}
r_{10}	{294.10, 295.04, 296.03, 834.0, 944.01}

(a)

Diagnosis code	Description
294.10	Dementia in conditions classified elsewhere without behavioral disturbance
295.04	Simple type schizophrenia, chronic with acute exacerbation
296.00	Bipolar I disorder, single manic episode, unspecified
296.01	Bipolar I disorder, single manic episode, mild
296.02	Bipolar I disorder, single manic episode, moderate
296.03	Bipolar I disorder, single manic episode, severe, without mention of psychotic behavior
401.0	Malignant essential hypertension
404.00	Hypertensive heart and chronic kidney disease, malignant, without heart failure and with chronic kidney disease stage I through stage IV, or unspecified
480.1	Pneumonia due to respiratory syncytial virus
692.71	Sunburn
695.10	Erythema multiform, unspecified
834.0	Closed dislocation of finger, unspecified part
944.01	Burn of unspecified degree of single digit (finger (nail) other than thumb

(b)

with r_1 , in Table 7.1a. This is because the set of ICD codes {296.01, 834.0} appears in *no other record* of the data in Table 7.1a.

Identity disclosure attacks, based on diagnosis codes, were first studied by Loukides et al. in [10], and they led to the development of several data anonymization methods (e.g., [11, 12, 14, 15, 18]). At an abstract level, these methods transform

a dataset to ensure that an attacker cannot associate an individual with fewer than k records, where k is a parameter that is specified by data owners. *In other words, these methods ensure that the probability of performing identity disclosure, based on diagnosis codes, will not exceed $1/k$.*

The transformation of diagnosis codes is performed either using *generalization* (i.e., by replacing diagnosis codes with more general, but semantically consistent, terms) and *suppression* (i.e., by deleting selected diagnosis codes) [4, 12], or using *disassociation* (i.e., by segmenting a record into subrecords that contain non-transformed diagnosis codes) [15].

7.1.2 Utility-Constrained Approach

In this chapter, we focus on anonymizing transaction datasets in a way that prevents identity disclosure and ensures that the shared data remain useful for intended studies. To clarify the relation between data anonymization and utility-preservation, assume that a healthcare institution wants to anonymize the dataset in Table 7.1a, so that: (i) the probability of identity disclosure, based on all sets of two diagnosis codes, does not exceed $1/3$, and (ii) the findings of studies u_1 to u_5 (see Table 7.2a), which require accurately counting the number of patients diagnosed with any combination of codes in them, are preserved.

Requirement (i) is posed by the need to comply with data sharing policies, while satisfying requirement (ii) will ensure that the data can be meaningfully analyzed by collaborating researchers, who perform the studies u_1 to u_5 . The anonymized dataset in Table 7.2b is produced using generalization. For example, the diagnosis codes 294.10, 295.04, 296.03, 401.0, 404.00, 480.1, and 944.01 are all replaced by the *generalized term* (294.10, 295.04, 296.03, 401.0, 404.00, 480.1, 944.01). This term is interpreted as *any combination (subset) of these codes*.

Generalization prevents an attacker from identifying a patient, based on knowledge of the patient's diagnosis codes, as the attacker cannot distinguish between the combination of diagnosis codes in the generalized term [4, 12]. Observe that the dataset in Table 7.2b satisfies requirement (i), because each combination of 2 diagnosis codes appears in at least 3 records. For instance, an attacker who knows that an individual is diagnosed with *Bipolar I disorder, single manic episode, mild* and *Closed dislocation of finger, unspecified part* (i.e., the codes 296.01 and 834.0) cannot uniquely associate a patient with r_1 in Table 7.2b. This is because 7 other records (i.e., the records $r_2, r_4, r_6, r_7, r_8, r_9$, and r_{10}) are associated with these two codes. However, requirement (ii) is *not* satisfied, as none of the studies u_1 to u_5 can be performed accurately using the anonymized dataset in Table 7.2b.

Consider, for example, the study $u_4 = \{480.1\}$. The number of patients associated with 480.1 (i.e., the *support* of this code) is not the same in the dataset of Table 7.1a and in its anonymized counterpart in Table 7.2b. Consequently, the study u_4 can no longer be performed accurately. Sharing data that affects the findings of intended studies can have serious consequences. For instance, such data may lead to

Table 7.2 (a) Utility requirements, and (b) anonymized dataset that *does not* satisfy the utility requirements (reprinted from [15], with permission from Elsevier)

ID	Utility constraints
u_1	{294.10, 295.04, 296.00, 296.01, 296.02, 296.03}
u_2	{692.71, 695.10}
u_3	{401.0, 404.00}
u_4	{480.1}
u_5	{834.0, 944.01}

ID	Records
r_1	(294.10, 295.04, 296.03, 401.0, 404.00, 480.1, 944.01), (296.00, 296.01, 296.02, 692.71, 695.10, 834.0)
r_2	(294.10, 295.04, 296.03, 401.0, 404.00, 480.1, 944.01), (296.00, 296.01, 296.02, 692.71, 695.10, 834.0)
r_3	(294.10, 295.04, 296.03, 401.0, 404.00, 480.1, 944.01)
r_4	(294.10, 295.04, 296.03, 401.0, 404.00, 480.1, 944.01), (296.00, 296.01, 296.02, 692.71, 695.10, 834.0)
r_5	(294.10, 295.04, 296.03, 401.0, 404.00, 480.1, 944.01)
r_6	(296.00, 296.01, 296.02, 692.71, 695.10, 834.0)
r_7	(294.10, 295.04, 296.03, 401.0, 404.00, 480.1, 944.01), (296.00, 296.01, 296.02, 692.71, 695.10, 834.0)
r_8	(296.00, 296.01, 296.02, 692.71, 695.10, 834.0)
r_9	(296.00, 296.01, 296.02, 692.71, 695.10, 834.0)
r_{10}	(294.10, 295.04, 296.03, 401.0, 404.00, 480.1, 944.01), (296.00, 296.01, 296.02, 692.71, 695.10, 834.0)

(b) Anonymized dataset

results that are difficult to interpret or to the discovery of spurious knowledge, when records contain clinical profiles (e.g., demographics and/or genomic sequences, in addition to diagnosis codes) [10, 18]. Intuitively, an intended study can be performed fairly accurately when the number of patients associated to the diagnosis codes of the study does not increase or decrease significantly, as a result of anonymization.

To anonymize data in a way that supports intended studies, a *utility-constrained* approach can be adopted [12]. In this approach, data owners specify sets of diagnosis codes that model analytic tasks that the published data are intended for. Each such set is termed a *utility constraint* and is used to control the way that diagnosis codes will be transformed. For instance, each of the studies u_1 to u_5 , mentioned above, corresponds to a utility constraint, as shown in Table 7.2a. The specified utility constraints are taken into account during data transformation to prevent their violation.⁵ Consider, for example, the anonymized dataset in Table 7.3, which is

⁵As we explain later in the chapter, some utility constraints may not be satisfied in the transformed data, due to the hardness of the problem.

Table 7.3 Anonymized counterpart of the dataset in Table 7.1a, using the utility-constrained approach (reprinted from [15], with permission from Elsevier)

Cluster	$ P_1 = 5$	Record chunks		Item chunk	
		C_1	C_2		
Cluster	r_1	{296.00, 296.01, 296.02}			
	r_2	{296.00, 296.01, 296.02}	{692.71, 695.10}		
	r_3	{296.00, 296.02}	{692.71, 695.10}	401.0, 834.0, 944.01	
	r_4	{296.00, 296.01}	{692.71}		
	r_5	{296.00, 296.01, 296.02}	{692.71, 695.10}		
Cluster	$ P_2 = 5$	Record chunk	Item chunk		
		C_1	C_T		
Cluster	r_6	{296.03, 295.04}			
	r_7	{294.10, 296.03}			
	r_8	{294.10, 295.04, 296.03}	404.00, 480.1, 834.0, 944.01		
	r_9	{294.10, 295.04}			
	r_{10}	{294.10, 295.04, 296.03}			

produced based on the utility-constrained approach. This dataset is comprised of two clusters, and each record of the dataset is comprised of a number of subrecords, called *chunks*. Notice that the anonymized dataset contains diagnosis codes that are not transformed, but the associations between chunks are “blurred”. For example, the set of diagnosis codes {401.0, 834.0} may be associated with any of the records r_1 to r_5 in the first cluster. The anonymized dataset achieves the same amount of privacy as the one in Table 7.2b, and it additionally satisfies the utility constraint $u_4 = \{480.1\}$ in Table 7.2a*. This is because the number of patients associated with 480.1 (i.e., the support of this code) is the same as in the dataset of Table 7.1a. Thus, the study corresponding to u_4 can be performed with no accuracy loss.

7.1.3 Chapter Organization

In the remainder of the chapter, we discuss the utility-constrained approach in detail. Specifically, Sect. 7.2 presents some concepts that are necessary to explain the utility-constrained approach. Section 7.3 explains the operations of generalization and disassociation that are used to transform diagnosis codes, in order to preserve privacy. Section 7.4 elaborates on the specification of utility constraints. Section 7.5 presents state-of-the-art algorithms that employ generalization and disassociation to anonymize data, and Sec. 7.6 provides a review of promising directions for future work. Last, Sect. 7.7 concludes the chapter.

7.2 Preliminaries

In this section, we define the type of data we consider, as well as how identity disclosure can be performed, using such data. Subsequently, we present a privacy principle that can be used to thwart identity disclosure.

We assume a dataset, D , which is comprised of $|D|$ records. We will refer to D as the *original dataset*. Each record in dataset D corresponds to a distinct patient, and it contains the set of all diagnosis codes that are associated with this patient. The number of records of D in which a set of diagnosis codes appears will be referred to as the *support* of the set in D . All diagnosis codes are derived from a finite domain, denoted with T , and the diagnosis codes contained in D are derived from a subset of T , denoted with T_D . An example of an original dataset is shown in Table 7.1a. Each record in the dataset contains some diagnosis codes, and the domain T contains all diagnosis codes in Table 7.1b. Notice that, in contrast to the traditional attack models for relational data [9, 16], the methods following the utility-constrained approach do not distinguish between *sensitive* (unknown to the attacker) and *non-sensitive* diagnosis codes in a record. Instead, they assume that any diagnosis code may be used in identity disclosure attacks.

In the following, we formally describe how identity disclosure may be performed. We assume that an attacker knows up to m diagnosis codes of an individual, whose record r is contained in D . Clearly, m can be any integer from 0 to m_{max} , where m_{max} is the maximum number of diagnosis codes that appear in a record of D . The case in which an attacker has no background knowledge is modeled by setting m to 0, whereas knowledge of all diagnosis codes about an individual is modeled by setting m to m_{max} . Also, there may be multiple attackers, each of whom knows a (not necessarily distinct) set of up to m diagnosis codes. Based on this knowledge, an identified patient can be associated with their record r , by an attacker. As we will discuss later, the parameter m is set by data owners, according to data sharing policies. To thwart this threat, the privacy model of k^m -anonymity [23] may be employed. k^m -anonymity is a conditional form of k -anonymity [20, 21], which ensures that an attacker, with partial knowledge of a record r , will not be able to distinguish r from $k-1$ other records in the published dataset. In other words, the probability that the attacker performs identity disclosure is upper bounded by $1/k$.

Definition 7.1 (k^m -anonymity [23]). Given data owner-specified parameters k and m , an anonymized dataset D^A satisfies k^m -anonymity, if no attacker with background knowledge of up to m diagnosis codes about an individual, represented in D^A , can associate the individual with fewer than k candidate records in D^A .

We now introduce two transformations that are used to define the privacy guarantee against identity disclosure. The anonymization transformation, denoted with \mathcal{A} , takes as input the original dataset D and produces an anonymized dataset D^A . Note that this transformation is independent of the operation that is applied to diagnosis codes. The inverse transformation, denoted with \mathcal{A}_I , takes as input the anonymized dataset D^A and outputs all possible (non-anonymized) datasets that could produce

D^A , i.e., $\mathcal{A}_I(D^A) = \{D' \mid D^A = \mathcal{A}(D)\}$. Obviously, the original dataset D is one of the datasets in $\mathcal{A}_I(\mathcal{A}(D))$. Thus, k^m -anonymity (Definition 7.1) can be satisfied, by enforcing the following privacy guarantee (adapted from [22]).

Guarantee 1. Given an anonymized dataset D^A and a set of up to m diagnosis codes, applying $\mathcal{A}_I(D^A)$, will produce at least one dataset $D' \in \mathcal{A}_I(D^A)$, for which there are at least k records that contain all diagnosis codes in the set.

Intuitively, an attacker, who knows any set of up to m diagnosis codes about an individual, will have to consider at least k candidate records in a possible original dataset. For example, assume that an attacker knows the set of codes $\{401.0, 834.0\}$ about an individual. The dataset in Table 7.2b, which was produced by generalization, satisfies 3^2 -anonymity, as the attacker cannot associate an individual with fewer than 3 records (i.e., they have to consider the records r_1, r_2, r_4, r_7 , and r_{10}). Similarly, the dataset in Table 7.3, which was produced by disassociation, also satisfies 3^2 -anonymity, as explained in the Introduction.

7.3 Generalization and Disassociation

In this section, we explain the operations of generalization and disassociation, which are used to prevent identity disclosure, in the utility-constrained approach. We first discuss generalization. As mentioned in the beginning of the chapter, this operation replaces a diagnosis code with a more abstract, but semantically consistent, *generalized term*. Generalization was first applied to solve the problem of anonymizing demographic attributes by Sweeney [21]. Suppression, also mentioned earlier, can be thought of as the special case of generalization where diagnosis codes are replaced by the most general concept, *Any* (usually denoted by *), which is interpreted as any diagnosis code in the domain T . Furthermore, generalization typically incurs a lower amount of information loss than suppression. Thus, we will not discuss suppression separately, in the remainder of this section.

Diagnosis codes can be generalized, using *global* or *local* models. Global models require generalizing all instances (i.e., occurrences) of a diagnosis code, in the original dataset, in the same way. On the contrary, local models do not impose this requirement. The dataset shown in Table 7.2b, for example, has been produced by applying a global generalization model to the dataset of Table 7.1a. Note that all instances of the diagnosis codes 294.10, 295.04, 296.03, 401.0, 404.00, 480.1, and 944.01 have been replaced by (294.10, 295.04, 296.03, 401.0, 404.00, 480.1, 944.01), which implies that a patient could have been diagnosed with any combination of these codes.

While local models are known to reduce information loss, they may lead to the construction of datasets that are difficult to be used in practice. This is because data mining algorithms and analysis tools cannot work effectively on these datasets [2]. Thus, most works have adopted a global model to generalize diagnosis codes. This model is referred to as *set-based generalization*, and it is defined as follows.

Table 7.4 Mappings between diagnosis codes and generalized terms, created by set-based generalization

Diagnosis code	Generalized term
294.10	(294.10, 295.04, 296.03, 401.0, 404.00, 480.1, 944.01)
295.04	
296.03	
401.0	
404.00	
480.1	
944.01	
296.00	(296.00, 296.01, 296.02, 692.71, 695.10, 834.0)
296.01	
296.02	
692.71	
695.10	
834.0	

Definition 7.2 (Set-based generalization [12]). Given a dataset D and a set of diagnosis codes T_D that contains all diagnosis codes in D , a *set-based generalization* is a partition \tilde{I} of T_D .⁶ The partition \tilde{I} is constructed by mapping each diagnosis code i in T_D to a *generalized term* \tilde{i} that contains i .

Set-based generalization requires replacing each diagnosis code in T_D with a unique generalized term. For example, the anonymized dataset in Table 7.2b has been produced by applying set-based generalization. Each code in the domain of codes in Table 7.1b has been replaced by a generalized term, and the two generalized terms form a partition of the domain. The mapping between the diagnosis codes and the generalized terms is shown in Table 7.4.

The application of set-based generalization can prevent identity disclosure, as it can lead to an increase in the number of records that can be associated with an individual. This is because *a generalized term appears in at least as many records as any of the diagnosis codes it replaces*. For instance, the generalized term (294.10, 295.04, 296.03, 401.0, 404.00, 480.1, 944.01) appears in 7 records in Table 7.2b, while all diagnosis codes in this generalized term appear in no more than 4 records in Table 7.1a. A slightly different model to generalize diagnosis codes, which are represented as ICD codes,⁷ was used in [11]. The difference is that, in [11], the generalized terms must be nodes in a hierarchy that is specified by data owners. The hierarchy is a tree structure, where leaves represent diagnosis codes

⁶A partition of a set is defined as a set of non-empty subsets, such that every element in the set is in exactly one of these subsets.

⁷ICD is a standard coding scheme for representing diagnosis codes. More details can be found at <http://www.who.int/classifications/icd/en/>.

and the root represents the most general concept (*Any*). If diagnoses are represented using ICD codes, then the standard ICD code hierarchy can be used (see [11]).

We now turn our attention to *disassociation*. This operation was proposed by Terrovitis et al. in [22] and was recently employed for protecting diagnosis codes by Loukides et al. [15]. Disassociation partitions the records in the original dataset D into subrecords. The main principle behind disassociation is that combinations of diagnosis codes that appear “few” times in D can be concealed, as they will be scattered in different subrecords of the published dataset. The general concept of breaking associations between values has also been employed in [25], for relational data (e.g., data containing demographic attributes).

In the following, we illustrate disassociation through an example. Consider Table 7.3, which shows a disassociated dataset produced from the original dataset of Table 7.1a. The dataset in Table 7.3 is comprised of two clusters, P_1 and P_2 , which contain the records r_1 to r_5 and r_6 to r_{10} , respectively. Furthermore, the diagnosis codes in a cluster are divided into subsets, and each record in the cluster is split into subrecords, according to these subsets. For example, the diagnosis codes in P_1 are divided into subsets $T_1 = \{296.00, 296.01, 296.02\}$, $T_2 = \{692.71, 695.10\}$, and $T_T = \{834.0, 401.0, 944.01\}$, according to which r_1 is split into three subrecords: $\{296.00, 296.01, 296.02\}$, an empty subrecord $\{\}$, and $\{834.0, 944.01\}$.

The collection of all (possibly empty) subrecords of different records that correspond to the same subset of diagnosis codes is called a *chunk*. For instance, the subrecord $\{296.00, 296.01, 296.02\}$ of r_1 goes into chunk C_1 , the empty subrecord goes into chunk C_2 , and the subrecord $\{834.0, 944.01\}$ goes into chunk C_T . In contrast to C_1 and C_2 which are *record chunks*, C_T is a special, *item chunk*, containing a single set of diagnosis codes. In our example, C_T contains the set $\{834.0, 401.0, 944.01\}$, which represents the subrecords from all r_1 to r_5 containing these codes. Thus, the number of times each diagnosis code in C_T appears in the original dataset is completely hidden from an attacker, who can only assume that this number ranges from 1 to $|P_i|$, where $|P_i|$ is the number of records in P_i .

Note that, different from generalization and suppression, disassociation retains all diagnosis codes of the published dataset in their original form. Thus, no information is lost within each record chunk, which is helpful when a chunk contains diagnosis codes that are analyzed together. This property makes the use of disassociation particularly beneficial in the utility-constrained approach, as we will discuss later. However, in order to preserve privacy, associations between diagnosis codes that are contained in different chunks are lost. In fact, the original dataset may contain any record that could be *reconstructed* by a combination of subrecords from the different chunks plus *any subset of diagnosis codes* from C_T .

For example, $\{296.00, 296.01, 834.0, 944.01\}$ in Table 7.5 is a reconstructed record, which is created by taking $\{296.00, 296.01\}$ from C_1 , the empty subrecord $\{\}$ from C_2 , and $\{834.0, 944.01\}$ from C_T . Observe that this record does not appear in the original dataset of Table 7.1a. In other words, the disassociated dataset can be interpreted as *any possible original dataset*, as explained in Guarantee 1. This way, the original dataset D is *hidden*, among all possible datasets that can be reconstructed from a disassociated dataset.

Table 7.5 A possible dataset reconstructed from the dataset of Table 7.3 (reprinted from [15], with permission from Elsevier)

ID	Records
r_1	{296.00, 296.01, 834.0, 944.01}
r_2	{296.02, 296.01, 692.71, 834.0}
r_3	{296.00, 296.01, 296.02, 692.71, 695.10, 834.0}
r_4	{296.00, 296.02, 692.71, 695.10}
r_5	{296.00, 296.01, 296.02, 692.71, 695.10, 401.0}
r_6	{296.02, 295.04, 480.1}
r_7	{294.10, 296.02, 404.00, 834.0, 944.01}
r_8	{294.10, 295.04, 296.02, 480.1, 834.0}
r_9	{294.10, 295.04, 404.00, 834.0}
r_{10}	{294.10, 295.04, 296.02, 834.0, 944.01}

Reconstructed datasets can permit accurate analysis, as (i) they retain most of the statistical properties of the original dataset, and (ii) they are straightforward to analyze (e.g., using off-the-shelf tools), in contrast to generalized datasets that require special handling (e.g., interpreting a generalized term as an original diagnosis code, with a certain probability) [15]. However, it is important to note that a reconstructed dataset may contain associations between diagnosis codes that do not exist in the original dataset. For example, the combination {296.02, 480.1} appears in the record r_6 of the reconstructed dataset in Table 7.5, but it does not appear in the original dataset. The extent to which such associations affect utility can be controlled, and it is minimal in practice, as discussed in [22].

7.4 Specification of Utility Constraints

In this section, we explain how utility constraints can be specified by data owners. First, we discuss the notion of utility constraints and their satisfaction. Subsequently, we review different types (instantiations) of utility constraints that have been proven useful in the context of medical applications.

7.4.1 Defining and Satisfying Utility Constraints

Both generalization and disassociation incur information loss, which must be controlled to ensure that anonymized data can support intended studies. To achieve this, a utility-constrained approach has been proposed in [12]. In addition to minimizing information loss, as most approaches do [6, 22, 27], the utility-constrained approach requires that a set of utility requirements, specified by data owners, are satisfied in the released data. These requirements are called *utility constraints*, and the set of the specified utility constraints is termed *utility policy*. The following definitions explain the concept of utility constraint and utility policy, and they are adapted from [12].

Definition 7.3 (Utility Constraint and Utility Policy [12]). A *utility constraint* u is a set diagnosis codes, specified by data owners. The set of all utility constraints is a partition of T_D , which is called *utility policy* and is denoted with \mathcal{U} .

As an example of a utility constraint, consider $u_2 = \{692.71, 695.10\}$ in Table 7.2a. This utility constraint is comprised of 2 diagnosis codes and is contained in the utility policy $\{u_1, u_2, u_3, u_4, u_5\}$. Note that the utility constraints are non-overlapping and that *each diagnosis code is contained in exactly one utility constraint*. Thus, a utility policy is a partition of the domain of diagnosis codes of a given dataset.

By specifying a utility constraint u , we aim to ensure that *the number of patients diagnosed with any diagnosis in u will be the same before and after anonymization*. If this holds, the anonymized dataset will be as useful as the original dataset, for the study corresponding to u . Furthermore, it is easy to see that the usefulness of the anonymized data for the study will be lower, if generalization or disassociation incurs a significant change in the number of patients diagnosed with u . This intuition leads to the following measure, called *Matching Relative Error* (MRE) [15].

Definition 7.4 (Matching Relative Error [15]). Let u be a utility constraint in a utility policy \mathcal{U} , and M_A and M_O be functions, which return the number of records that *match* u in the anonymized dataset D' and in the original dataset D , respectively. The MRE for u is computed as:

$$MRE(u) = \frac{M_O(u) - M_A(u)}{M_O(u)}$$

Thus, a zero MRE implies that an anonymized dataset can support u to the same extend as the original dataset does, and MRE scores close to zero are preferred. In addition, MRE can take positive (respectively, negative) values, when the number of records that match u in the anonymized dataset is smaller (respectively, larger) than the corresponding number in the original dataset. The sign of MRE helps distinguishing between these two cases, which may be useful for specific studies (e.g., when underestimating the number of patients is not as harmful as overestimating it). Note also that MRE is independent of the data transformation operation, hence it can be applied to generalized or disassociated datasets. For a generalized or reconstructed dataset, a record matches u if it contains at least one of the diagnosis codes in u . For a disassociated dataset, the definition of a match is slightly different. Specifically, given a record r in a cluster P of a disassociated dataset, we say that r matches a utility constraint u , if at least one of the diagnosis codes in u is contained in: (i) the record chunk or r , or (ii) in the item chunk of P .

For example, the utility constraint u_2 in Table 7.2a matches 4 records (namely r_2, r_3, r_4, r_5) in the original dataset of Table 7.1a. However, u_2 matches 8 records in the generalized dataset, shown in Table 7.2b, because at least one of the codes in $\{692.71, 695.10\}$ appears in 8 records. Thus, in this case, $MRE(u_2) = -1$. This implies that the number of records that match u_2 in the generalized dataset is 100 % (i.e., 2 times) larger than that in the original dataset. Now consider u_2 in

the disassociated dataset of Table 7.5, in which u_2 matches 4 records (namely r_2 to r_5). In this case, $MRE(u_2) = 0$. Since the MRE for u_2 in the dataset of Table 7.5 is lower than that for the dataset of Table 7.2b, the former dataset is more useful for u_2 .

When the MRE for a utility constraint is zero, we say that this constraint is *satisfied*. A *utility policy* is *satisfied* when all utility constraints in it are satisfied. Given an anonymized dataset and a utility policy, one may explicitly compute M_O and M_A , for each utility constraint in the utility policy, to find out whether the utility policy is satisfied. However, to produce a generalized (respectively, disassociated) dataset that satisfies a utility policy, it is necessary to study how generalization (respectively, disassociation) affects M_A . The following property explains this for the case of data generalization.

Property 7.1. A utility policy \mathcal{U} is *satisfied* for a generalization \tilde{I} , if and only if, for each generalized term \tilde{i} in \tilde{I} , all diagnosis codes that are mapped to \tilde{i} are contained in a single utility constraint in \mathcal{U} .

That is, \tilde{I} should not contain a diagnosis code whose corresponding generalized term “spans” utility constraints. This property holds, because all occurrences of a diagnosis code are mapped to the same generalized term, in all records of the dataset. Thus, a good strategy to produce a generalized dataset that satisfies the utility policy is to generalize diagnosis codes, contained in the same utility constraint.

In the following, we provide a property for disassociation. This property is applied to the clusters of a disassociated dataset, because the matching of a record to a utility constraint is performed for a cluster, as explained above.

Property 7.2. Given a utility constraint u and a disassociated dataset, comprised of clusters C_1, \dots, C_r , the utility constraint is *satisfied* in the disassociated dataset, if the diagnosis codes of u are contained in at most one record chunk, for each cluster C_i , where $i \in [1, r]$. A utility policy \mathcal{U} is *satisfied* for a disassociated dataset, if each utility constraint u in \mathcal{U} is satisfied.

Thus, the diagnosis codes of a utility constraint should not “span” multiple record chunks of a cluster. Otherwise, diagnosis codes that appear in the same record of the original dataset can end up in different records in the disassociated dataset. This, in turn, may lead to the violation of the utility constraint. For example, assume that a utility constraint contains two diagnosis codes, which appear in a single record of the original dataset. The utility constraint will not be satisfied when these codes end up in two different record chunks and each chunk belongs to a different record of the disassociated dataset.

Based on Properties 7.1 and 7.2, we make the following observations, for a utility constraint u . These observations are important to select an appropriate anonymization algorithm in practice, and they have not been discussed so far in the research literature, to the best of our knowledge.

Observation 1: The number of individuals that are associated with *any subset* of diagnosis codes in u is not affected by disassociation, when u is satisfied. That is, any utility constraint that is more fine-grained than u is also satisfied. This does

not necessarily hold when generalization is used. In other words, *generalization should be used when a study does not distinguish between diagnosis codes in u, whereas disassociation should be used when such distinction is important to be made.*

Observation 2: A utility constraint u that contains many diagnosis codes is generally more difficult to be satisfied by disassociation, whereas the opposite holds for generalization. For example, a utility constraint that contains all diagnosis codes in the domain T will always be satisfied by generalization, whereas creating record chunks that end up in different records in a disassociated dataset may lead to the violation of the utility constraint. Thus, *the use of disassociation should be preferred when studies involve a small number of diagnosis codes.*

7.4.2 Types of Utility Constraints for ICD Codes

The concept of utility constraint is fairly general in the sense that any set of diagnosis codes can form a utility constraint. Consequently, utility constraints can be formed by grouping diagnosis codes represented using different medical terminologies, such as ICD or SNOMED CT.⁸ However, most works on anonymizing diagnosis codes (see [5] for a survey) focus on utility constraints formed of ICD codes, which is also the focus of this chapter. In the following, we will present three types of utility constraints, which are based on: (i) the ICD hierarchy, (ii) the similarity between ICD codes, or (iii) the frequency of ICD codes in the original dataset. These types of constraints can be used to model the requirements of various medical studies, as explained in [11].

Hierarchy-Based Utility Constraints These constraints model semantic relationships between ICD codes that are based on the ICD hierarchy. This hierarchy has 5-digit ICD codes as leaves and 3-digit ICD codes, *Sections* and *Chapters*, as internal nodes. The root of this hierarchy is the most general concept *Any*. Sections and Chapters model aggregate concepts, and Chapters correspond to more general concepts than Sections. Thus, a utility constraint can contain all 5-digit ICD codes (leaves) that have a common ancestor (i.e., 3-digit ICD code, *Section*, or *Chapter*). The intuition behind hierarchy-based constraints is that a study often involves different forms of the same disease (e.g., different forms of Schizophrenic disorders) or classes of related diseases according to the ICD hierarchy (e.g., Psychoses).

For example, consider a utility constraint u for *Schizophrenic disorders*. The 5-digit ICD codes in u are of the form 295.xy, where $x = \{0, \dots, 9\}$ and $y = \{0, \dots, 5\}$, and they have the 3-digit ICD code 295 as their common ancestor. The utility constraint u is shown in the first row of Table 7.6. By forming a different utility constraint, for each 3-digit ICD code in the hierarchy (e.g., 296, 297, etc.), we construct a *level 1*, hierarchy-based policy. Alternatively, the common ancestor

⁸<http://www.ihtsdo.org/snomed-ct/>.

Table 7.6 Examples of hierarchy-based utility constraints

Utility policy	ICD codes in utility constraint
<i>level 1</i>	{295.00, 295.01, ..., 295.95}
<i>level 2</i>	{295.00, 295.01, ..., 295.95, 296.00, ..., 299.91}
<i>level 3</i>	{290.10, 295.00, 295.01, ..., 295.95, 296.00, ..., 299.91, ..., 299.91, ..., 319}

of the codes in the utility constraint u may be a *Section*. For example, u is comprised of *Psychoses*, whose common ancestor is 295–299, in the second row of Table 7.6. In this case, u will be contained in a *level 2*, hierarchy-based policy. In another case, the common ancestor for the codes in u may be a *Chapter*. For example, u may correspond to *Mental disorders* that have 290–319 as their common ancestor (see the last row of Table 7.6). In this case, u is contained in a *level 3*, hierarchy-based policy.

Similarity-Based Utility Constraints Similarity-based utility constraints model semantic relationships between ICD codes, which are not necessarily formed using a hierarchy. Thus, they are more general than hierarchy-based utility constraints.

For example, they can be used to model the requirements of studies involving certain forms of a disorder (i.e., only some of the 5-digit ICD codes whose common ancestor is a 3-digit ICD code) or broader groups of ICD codes (e.g., some 3-digit ICD codes which do not have the same common ancestor). In addition, different forms of similarity (e.g., other taxonomies or distance measures) can be used. In [11], for instance, similarity-based utility constraints that contain the same number of sibling 5-digit ICD codes in the hierarchy have been considered. As an example, a similarity-based utility constraint can contain 5 sibling ICD codes: 295.00, 295.01, 295.02, 295.03, and 295.04. These codes are only some of the 5-digit ICD codes that have the 3-digit ICD code 295 as their common ancestor in the ICD hierarchy.

Frequency-Based Utility Constraints Frequency-based utility constraints model *frequent itemsets* (i.e., sets of diagnosis codes that appear in at least a certain percentage of the records of the original dataset). For example, a utility constraint can contain all diagnosis codes that appear in at least 5 % percent of records, and a utility policy may be formed of all such utility constraints. The intuition behind frequency-based utility constraints is that studies often involve codes that appear relatively frequent in the original dataset.

7.5 Utility-Constrained Anonymization Algorithms

In this section, we provide an overview of the *Clustering-based Anonymizer* (CBA) [11] and the *DISassociation* (DIS) [15] algorithm. These algorithms anonymize data following the utility-constrained approach, and they are the current state-of-the-art

in terms of their effectiveness to preserve data utility. Subsequently, we present a concrete example that illustrates the operation of the CBA and DIS algorithms, as well as the differences between them.

7.5.1 Clustering-Based Anonymizer (CBA)

The CBA algorithm aims to enforce k^m -anonymity to an original dataset by applying generalization to diagnosis codes, so that the utility policy is satisfied and minimal information loss is incurred.⁹ The pseudocode of CBA is provided in Algorithm 7.1.

Algorithm 7.1 CLUSTERING-BASED ANONYMIZER (CBA).

Input : Original dataset D , utility policy \mathcal{U} , parameters k and m

Output : Generalized dataset D^A

```

1  $D^A \leftarrow D$ 
2  $PQ \leftarrow$  all sets of up to  $m$  diagnosis codes with support lower than  $k$  in  $D$ 
3 while  $PQ$  is not empty do
4    $p \leftarrow$  set of diagnosis code with largest support
5   for each diagnosis code  $t$  in  $p$  do
6     if  $t$  has been generalized before then
7       | Update  $p$  by replacing  $t$  with its generalized term
8     end
9   end
10  while the support of  $p$  in  $D^A$  is in  $(0, k)$  do
11    Find diagnosis codes  $t$  and  $t'$ , such that (i)  $t$  and  $t'$  are in the same utility constraint, and
12      (ii)  $(t, t')$  has the lowest information loss
13    if such codes can be found then
14      | Create the generalized term  $(t, t')$ 
15      | Update  $p$  and  $D^A$  by replacing  $t$  and  $t'$  with  $(t, t')$ 
16    end
17    else
18      | while the support of  $p$  in  $D^A$  is in  $(0, k)$  do
19        |   Delete the diagnosis code in  $p$  that has the lowest support in  $D^A$ 
20      end
21    end
22  Remove  $p$  from  $PQ$ 
23 end
24 return  $D^A$ 

```

This algorithm starts by initializing the dataset to be anonymized, denoted with D^A , to the original dataset D . Then, in step 2, a priority queue PQ is populated with

⁹This algorithm can also be used to apply a privacy model called *privacy-constrained anonymity* (see [11] for details). However, in this chapter, we focus on k^m -anonymity, as it is enforced by DIS.

all sets of diagnosis codes that need protection (i.e., those with up to m diagnosis codes and support lower than k in D). The priority queue keeps these sets sorted in decreasing order of their support in D .

In steps 4–23, CBA considers each of the elements in PQ . Specifically, CBA retrieves p , the top-most element of PQ , and updates the diagnosis code contained in it (steps 4–8). The update reflects generalizations that may have occurred in previous iterations, and it is necessary to decide if the element p requires protection. Thus, CBA iterates over each diagnosis code in p and replaces the code with its corresponding generalized term, if the code has been generalized before.

Subsequently, in steps 10–21, CBA protects p , if its support is lower than k in D^A . To achieve this, the algorithm attempts to find a pair of diagnosis codes t and t' that can be generalized together, according to the utility policy and in a way that minimizes information loss (step 11). Information loss is measured by taking into account the support of the generalized term (t, t') , as well as the semantic distance between t and t' (see [11] for details). If such a pair is found, t and t' are generalized to (t, t') , while the set of diagnosis codes p and D^A are updated to reflect the generalization (steps 12–15). Otherwise, CBA iteratively deletes diagnosis codes from p , starting with the least supported code in D^A , until p becomes protected (steps 16–20). This allows CBA to delete no more ICD codes than what is required to protect p . After p is satisfied, it is removed from PQ (step 22), and the algorithm checks whether PQ contains another element that must be protected. When PQ becomes empty, CBA returns the anonymized dataset D^A (step 24).

Time Complexity The worst case time complexity of CBA is $O(2^{|T_D|} \cdot |D|)$, where $|T_D|$ is the number of distinct diagnosis codes in the original dataset D , and $|D|$ is the number of records in D . This is because there are at most $O(2^{|T_D|})$ sets of diagnosis codes that require protection and computing the support of each set takes $O(|D|)$ time. In practice, the number of distinct diagnosis codes that require protection is smaller, and CBA is fairly efficient. For instance, anonymizing a dataset containing 25 K records and 631 distinct diagnosis codes required less than 6 s [15].

7.5.2 *DISassociation Algorithm (DIS)*

The DIS algorithm has the same objective as CBA (i.e., to enforce k^m -anonymity, while satisfying the utility policy and incurring minimal information loss), but it uses disassociation instead of generalization. Thus, the operations it performs are fundamentally different from those of CBA.

A high-level description of the DIS algorithm is provided in Algorithm 7.2. As can be seen, DIS performs a horizontal partitioning of the dataset into disjoint clusters, and then it partitions each cluster vertically. The creation of clusters is performed by an algorithm called HORPART, while the vertical partitioning of the clusters is performed by an algorithm called VERPART. Subsequently, DIS

Algorithm 7.2 DISASSOCIATION (DIS) (reprinted from [15], with permission from Elsevier).

Input : Original dataset D , utility policy \mathcal{U} , parameters k and m
Output : Disassociated dataset D^A

```

1 Partition  $D$  horizontally into disjoint clusters by applying Algorithm HORPART
2 for each cluster  $P$  do
3   | Partition  $P$  vertically into chunks by applying Algorithm VERPART
4 end
5 Refine clusters
6 return  $D^A$ 
```

performs the refining operation, which aims to further reduce information loss. In the following, we present the partitioning algorithms and the refining operation in more detail.

We first discuss HORPART, which horizontally partitions the original dataset into clusters. This algorithm aims to create clusters that contain semantically similar diagnosis codes and no more than $maxClusterSize$ records, where $maxClusterSize$ is a parameter provided by the data owners. The intuition behind the operation of HORPART is that the records in such clusters can be partitioned vertically without incurring significant information loss.

The pseudocode of HORPART is provided in Algorithm 7.3. As can be seen, the algorithm first checks whether the dataset it is applied to, denoted with D , contains fewer than $maxClusterSize$ records. In this case, the dataset D is returned. Otherwise, HORPART splits the dataset into two parts, according to: (i) the *support* of diagnosis codes in the original dataset, and (ii) the participation of diagnosis codes in the utility policy. The first part of D is denoted with D_1 and contains all records with a diagnosis code a . To select a , we distinguish two cases. In the first case (steps 4–6), all diagnosis codes in u have been considered in previous recursive calls of the algorithm. In this case, the diagnosis code with the largest support in D that has not been considered before is assigned to a , and the utility constraint that contains a is assigned to u . Otherwise, at least one diagnosis code, in the input utility constraint u , can be considered as a . Thus, a is assigned to the code in u that has the largest support in the original dataset (step 8). Subsequently, in steps 10–11, HORPART creates the two parts of the dataset; D_1 contains all records that contain a , and D_2 contains all the remaining records of D . This procedure is applied recursively, to each of the constructed parts (step 12), until they contain fewer than $maxClusterSize$ records. Diagnosis codes that have been previously used for partitioning are used only once.

Next, we discuss the VERPART algorithm, which vertically partitions a cluster into chunks. VERPART is essentially a greedy heuristic which tries to distribute non-protected combinations of codes into different chunks. By doing so, the algorithm breaks the associations between such codes, which helps creating a k^m -anonymous cluster. In addition, VERPART aims at satisfying the utility policy, by creating record chunks that contain as many diagnosis codes from the same utility constraint as possible. Note that this is in line with Property 7.2.

Algorithm 7.3 HORPART (reprinted from [15], with permission from Elsevier).

Input : Dataset D , utility policy \mathcal{U} , a utility constraint $u \in \mathcal{U}$ (initially empty)
Output : A Horizontal PARTitioning of D , i.e., a set of clusters
Param. : The maximum cluster size $maxClusterSize$

```

1 if  $|D| < maxClusterSize$  then
2   | return  $D$ 
3 end
4 if all diagnosis codes in  $u$  have been considered then
5   |  $a \leftarrow$  a diagnosis code that: (i) has the largest support in  $D$ , and (ii) has not been considered
       before  $u \leftarrow$  the utility constraint from  $\mathcal{U}$  that contains  $a$ 
6 end
7 else
8   |  $a \leftarrow$  a diagnosis code that: (i) is contained in  $u$ , (ii) has the largest support in  $D$ , and (iii) has
       not been considered before
9 end
10  $D_1 \leftarrow$  the set of all records of  $D$  that contain  $a$ 
11  $D_2 \leftarrow D - D_1$ 
12 return HORPART( $D_1, \mathcal{U}, u$ )  $\cup$  HORPART( $D_2, \mathcal{U}, \{\}$ )

```

The pseudocode of VERPART is provided in Algorithm 7.4. The algorithm starts by creating an item chunk that contains all diagnosis codes whose support in the cluster is lower than k . The support of a set of diagnosis codes cannot be larger than that of the diagnosis codes contained in it, hence such diagnosis codes could not be part of k^m -anonymous record chunks. Then, VERPART assigns all remaining diagnosis codes into a set T' , which is partitioned into groups (steps 2–3). Each group contains all diagnosis codes of a utility constraint, and these codes are sorted in decreasing order of their support in the cluster (step 4). Next, all groups of T' are sorted, in decreasing order with respect to the support of their first diagnosis code (step 5). This prioritizes the partitioning of clusters that contain at least one frequent code, and it was shown to be effective for preserving data utility [15].

Following that, in steps 6–18, the algorithm creates record chunks, by considering all diagnosis codes in T' . Specifically, each code t , which is not part of a record chunk, is used to create a new record chunk (step 8). All other codes that are contained in the same utility constraint as t are then considered, following the ordering of T' . These codes are added into the created record chunk, if the cluster remains k^m -anonymous (step 9). Clearly, the creation of large record chunks helps preserving the associations between diagnosis codes, while the selection of codes from the same utility constraint helps satisfying the utility policy. After that, in steps 11–13, VERPART considers each diagnosis code t in T' and finds: (i) the record chunk, $R(t)$, that contains t , and (ii) the first code, t' , that was added into $R(t)$. If t and t' belong to different utility constraints and some of the diagnosis codes in the utility constraint of t have not been added into $R(t)$, then t is removed from $R(t)$ (steps 14–15). This enables the algorithm to insert the code into another record chunk (along with the remaining codes of u) in a subsequent step. After all codes in T' are added into record chunks, the algorithm returns the disassociated cluster (step 19).

Algorithm 7.4 VERPART (reprinted from [15], with permission from Elsevier).

Input : A cluster P , utility policy \mathcal{U} , parameters k and m

Output : A k^m -anonymous VERTICAL PARTitioning of P

```

1 Create item chunk using the set of diagnosis codes in  $P$  that appear in fewer than  $k$  records of  $P$ 
2  $T' \leftarrow$  all diagnosis codes that are contained in  $P$  but not in the item chunk
3 Partition  $T'$  into groups, s.t. each group contains all diagnosis codes of a utility constraint in  $\mathcal{U}$ 
4 Sort the diagnosis codes of each group, in decreasing order of their support in  $P$ 
5 Sort the groups, in decreasing order with respect to the support of their first diagnosis code
6 repeat
7   for each diagnosis code  $t$  in  $T'$  that is not part of a record chunk do
8     Create an empty record chunk and add  $t$  into it
9     Add all diagnosis codes in  $T'$  that can be added into the record chunk without violating
       the  $k^m$ -anonymity of the cluster  $P$ , following the ordering of  $T'$ 
10    end
11   for each diagnosis code  $t$  in  $T'$  do
12     Let  $R(t)$  be the record chunk of  $t$ 
13     Let  $t'$  the diagnosis code that was first inserted into  $C(t)$ 
14     if  $t$  belongs to a utility constraint  $u$ , which is different from the constraint of  $t'$  and not all
       diagnosis codes of  $u$  are added to  $R(t)$  then
15       | Remove  $t$  from  $R(t)$ 
16     end
17   end
18 until all diagnosis codes in  $T'$  are contained in record chunks;
19 return disassociated cluster  $C$ 

```

Last, we explain the refining operation (see step 5 in the pseudocode of DIS). This operation aims at improving the utility of the disassociated dataset, without violating k^m -anonymity. To this end, we examine the diagnosis codes that reside in the item chunk of each cluster. Consider, for example, Table 7.3. The item chunk of the cluster P_1 contains the diagnosis codes 834.0 and 944.01, because the support of these codes in P_1 is 2 (i.e., lower than $k = 3$). For similar reasons, these diagnosis codes are also contained in the item chunk of P_2 . However, the support of these codes in both clusters P_1 and P_2 together is not small enough to violate privacy (i.e., the set {834.0, 944.01} appears as many times as the set {296.03, 294.10} which is in the record chunk of P_2). To deal with such situations, the DIS algorithm allows the creation of *joint clusters*, which share record chunks.

Given a set T^s of codes that appear in the item chunks of two or more clusters, a joint cluster can be created by (i) constructing one or more *shared chunks*, after projecting the original records of the initial clusters to T^s , and (ii) removing all diagnosis codes in T^s from the item chunks of the initial clusters. The shared chunks must be k^m -anonymous, to ensure that privacy is preserved, and contain a minimum number of subrecords, to help data utility, as explained in [22]. Table 7.7 shows a joint cluster, created by combining the clusters P_1 and P_2 of Table 7.3, when $T^s=\{834.0, 944.01\}$. Furthermore, larger joint clusters can be built by combining joint clusters.

Table 7.7 Disassociation with a shared chunk (reprinted from [15], with permission from Elsevier)

Record	Item	Shared
<i>P</i> ₁ cluster		
{296.00, 296.01, 296.02}		{834.0, 944.01}
{296.00, 296.01, 296.02}	{692.71, 695.10}	401.0
{296.00, 296.02}	{692.71, 695.10}	
{296.00, 296.01}	{692.71}	
{296.00, 296.01, 296.02}	{692.71, 695.10}	
<i>P</i> ₂ cluster		
{296.03, 295.04}		
{294.10, 296.03}	404.00,	{834.0, 944.01}
{294.10, 295.04, 296.03}	480.1	
{294.10, 295.04}		
{294.10, 295.04, 296.03}		{834.0, 944.01}

Time Complexity The worst-case time complexity of HORPART is $O(|D|^2)$, as it performs $O(|D|)$ splits and reorders $O(|D|)$ records in each split. The worst-case time complexity of VERPART is $O(|T'|!)$, as it examines $O(|T'|!)$ combinations of diagnosis codes, and that of the refining operation is $O(|D|^2)$. As noted in [15], the behavior of DIS is dominated by that of HORPART, because $|T|$ is small in practice (and it can also be controlled through varying *maxClusterSize*).

7.5.3 Comparing the CBA and DIS Algorithms

In this section, we provide an example that illustrates the application of the CBA and DIS algorithms to the dataset of Table 7.1a. The utility policy, shown in Table 7.2a, is used, and the parameters k and m are set to 3 and 2, respectively. Subsequently, we compare the outcome of these algorithms with respect to data utility.

We first illustrate the application of CBA. The algorithm begins by identifying all sets of up to 2 diagnosis codes that have a lower support than 3. There are totally 32 such sets of diagnosis codes, which are added into the priority queue, in decreasing order of their support (step 2). Table 7.8 shows some of these sets and their support. Then, CBA retrieves the top-most element of the priority queue, which is the set {296.00, 834.0}. None of the codes in the set has been generalized before, so no update takes place in steps 6–8. As the support of the set is lower than 3, the algorithm needs to apply generalization and/or suppression.

Specifically, CBA finds the pair 296.00 and 296.01 (step 11). These codes are semantically similar, hence incur low information loss, and they belong in the same utility constraint in Table 7.2a. Thus, the generalized term (296.00, 296.01) is created and these two codes are replaced by this term in the element of the

Table 7.8 Sets of diagnosis codes that are added into the priority queue of CBA, and their support

Support	Set of diagnosis codes
2	{296.00, 834.0}
2	{296.00, 944.01}
2	{296.00, 401.0}
2	{296.01, 944.01}
2	{296.01, 401.0}
...	
2	401.00
2	404.00
2	480.01
...	
1	{296.03, 404.00}
1	{401.0, 965.10}

priority queue, as well as in the original dataset (steps 12–14). After that, the top-most element of the priority queue becomes $\{(296.00, 296.01), 834.0\}$, which still requires protection as its support is lower than k . The algorithm proceeds in a similar fashion and eventually creates $(294.10, 295.04, 296.00, 296.01, 296.02, 296.03)$, whose support is 10 (i.e., larger than k).

Notice that the utility constraint u_1 in Table 7.1a is satisfied, but k^m -anonymity does not hold. This is because there are still sets of codes whose support is lower than k (e.g., $\{296.00, 401.0\}$). Therefore, CBA considers these sets and creates the generalized term $(401.0, 404.00)$. Still, u_2 is satisfied, but k^m -anonymity does not hold. At this point, there are no generalized terms that can be constructed, in a way that satisfies the remaining utility constraints. Thus, the algorithm employs suppression (i.e., it deletes codes), as shown in Table 7.9. For example, the code 480.1 has been suppressed, because it co-occurs with the generalized term $(401.0, 404.00)$ only in record r_6 , and generalizing 480.1 with another code would violate the utility constraint $u_4 = \{480.1\}$ in Table 7.2a. At this point, all sets of diagnosis codes have a support of at least k (i.e., the priority queue is empty), and CBA returns the anonymized dataset shown in Table 7.9 (step 24).

We now show how the DIS algorithm can be applied to the original dataset in Table 7.1a. We assume that the parameter $maxClusterSize$ is set to 6. The algorithm starts by using HORPART to create a set of disjoint clusters. As the size of the original dataset is larger than 6 and no diagnosis codes have been considered in previous calls (steps 1–6), the HORPART algorithm assigns a to 296.00 (step 5). This code is contained in the utility constraint u_1 of Table 7.2a and has the largest support in the original dataset. Then, the algorithm splits D into two parts, D_1 and D_2 . D_1 consists of the records containing 296.00 (i.e., r_1 to r_5), whereas D_2 contains the remaining records (steps 10–11). The next call of HORPART for D_1 is performed with the utility constraint u_1 as input, and HORPART tries to further partition D_1 , using the codes of this constraint. On the contrary, no utility constraints are given as input to HORPART, when it is applied to D_2 . As the size of both D_1 and D_2 is

Table 7.9 Anonymized dataset by CBA using the utility policy of Table 7.2a

ID	Records
r_1	(294.10, 295.04, 296.00, 296.01, 296.02, 296.03), 834.0, 944.01
r_2	(294.10, 295.04, 296.00, 296.01, 296.02, 296.03), (401.0, 404.00), 944.01, 692.71, 695.10
r_3	(294.10, 295.04, 296.00, 296.01, 296.02, 296.03), 692.71, 834.0, 695.10
r_4	(294.10, 295.04, 296.00, 296.01, 296.02, 296.03), (401.0, 404.00), 692.71
r_5	(294.10, 295.04, 296.00, 296.01, 296.02, 296.03), 692.71, 695.10
r_6	(294.10, 295.04, 296.00, 296.01, 296.02, 296.03), (401.0, 404.00), 480.1
r_7	(294.10, 295.04, 296.00, 296.01, 296.02, 296.03), 834.0, 944.01
r_8	(294.10, 295.04, 296.00, 296.01, 296.02, 296.03), 480.1
r_9	(294.10, 295.04, 296.00, 296.01, 296.02, 296.03), (401.0, 404.00)
r_{10}	(294.10, 295.04, 296.00, 296.01, 296.02, 296.03), 834.0, 944.01

The diagnosis codes appearing in grey have been suppressed (deleted)

lower than \maxClusterSize , HORPART creates two clusters; P_1 , which contains the first five records, and P_2 , which contains the remaining records. The two clusters amount to D_1 and D_2 , respectively.

Next, DIS applies the VERPART algorithm to each of the clusters P_1 and P_2 . VERPART starts by creating the item chunk containing the codes 401.0, 834.0 and 944.01, which appear in fewer than 3 records in P_1 (step 1). Thus, T' contains all other codes, and it is sorted according to the support of these codes in P_1 and their participation in a utility constraint (steps 2–4). Specifically, for the utility constraints of Table 7.2a, we distinguish two groups of codes; a group {296.00, 296.01, 296.02}, for the codes in u_1 , and another group {692.71, 695.10} for the codes in u_2 . Next, VERPART sorts each of these groups, in descending order of the support of its codes. Thus, the final ordering of T' is {296.00, 296.01, 296.02, 692.71, 695.10}.

Subsequently, VERPART constructs the record chunks of P_1 (steps 7–10), as follows. First, it selects 296.00, which has the largest support among all records of P_1 , and adds it into an initially empty record chunk. Then, it selects 296.01, the next code in T' , and adds it into the record chunk, as the record chunk remains 3^2 -anonymous. After that, it considers each of the remaining codes of T' , following the ordering of T' , and it adds 296.02 and 692.71 into the record chunk. The code 695.10 is not added, because the set {296.01, 695.10} appears in only two records of P_1 (i.e., r_2 and r_5), hence adding it into the chunk would violate 3^2 -anonymity.

After considering all codes in T' , VERPART checks whether the codes of a utility constraint are only partially added to the record chunk (steps 11–17). This is true for 692.71, which is separated from 695.10 (both these codes are contained in u_2). Hence, 692.71 is removed from the record chunk, so that it can be added to the chunk C_2 of P_1 along with 695.10. After that, the algorithm proceeds to creating the next record chunk. Then, DIS applies VerPart to the cluster P_2 , and creates the disassociated dataset shown in Table 7.3.

Following that, the algorithm refines the clusters of the dataset in Table 7.3 to reduce information loss. During this operation a shared chunk, shown in Table 7.10,

Table 7.10 Disassociated dataset with a shared chunk (reprinted from [15], with permission from Elsevier)

Record	Item	Shared
<i>P</i> ₁ cluster		
{296.00, 296.01, 296.02}		{834.0, 944.01}
{296.00, 296.01, 296.02}	{692.71, 695.10}	401.0
{296.00, 296.02}	{692.71, 695.10}	
{296.00, 296.01}	{692.71}	
{296.00, 296.01, 296.02}	{692.71, 695.10}	
<i>P</i> ₂ cluster		
{296.03, 295.04}		
{294.10, 296.03}		404.00, {834.0, 944.01}
{294.10, 295.04, 296.03}		480.1
{294.10, 295.04}		
{294.10, 295.04, 296.03}		{834.0, 944.01}

Table 7.11 The result of functions M_O and M_A , for CBA and for a reconstructed dataset, produced by DIS

Utility constraint	M_O	M_A for CBA	M_A for reconstructed
u_1	10	10	10
u_2	4	0	4
u_3	4	4	2
u_4	2	0	2
u_5	5	0	7

These functions are calculated for each utility constraint in Table 7.2a and used in the computation of the MRE measure

is created, as follows. First, DIS inspects the item chunks of P_1 and P_2 in Table 7.3, and it identifies that the codes 834.0 and 944.01 have support 2 in P_1 , as well as in P_2 . Since the support of these codes in both clusters together is $2 + 2 = 4 > k$, DIS reconstructs the parts $\{r_1, \dots, r_5\}$ and $\{r_6, \dots, r_{10}\}$ that contain these codes, and calls VERPART, which creates the shared chunk. Finally, the dataset created by DIS is used to generate the reconstructed dataset shown in Table 7.5.

In the following, we compare the anonymized dataset of CBA to the reconstructed dataset, created from the output of DIS, in terms of utility, using the MRE measure (see Sect. 7.4). The computation of MRE is based on the functions M_O and M_A , which are applied to each utility constraint in Table 7.2a. The results of the computation are shown in Table 7.11. Specifically, the latter figure shows the result of the functions M_A , for the dataset created by CBA, and for the reconstructed dataset, in the third and last column, respectively. Based on these functions, the MRE for each of the utility constraints is computed, as shown in Table 7.12. Observe that CBA outperforms DIS for u_3 , whereas it performs worse than DIS, in the case of u_2 , u_4 , and u_5 . The MRE scores for utility constraints in which an algorithm performs better are underlined.

Table 7.12 MRE scores for each utility constraint in Table 7.2a

Utility constraint	MRE for CBA	MRE for DIS
u_1	0	0
u_2	1	<u>0</u>
u_3	<u>0</u>	1
u_4	1	<u>0</u>
u_5	1	<u>-0.4</u>

Last, we compare CBA and DIS with respect to their ability to permit accurate aggregate query answering (i.e., to support studies requiring to find the number of patients associated with a certain combination of codes). The comparison is based on the INFORMS 2008 electronic medical record dataset [7], and it is part of the experimental evaluation of DIS reported in [15].

To measure aggregate query answering accuracy, we considered various workloads of queries, asking for sets of diagnosis codes that at least δ % of all patients have. In other words, the queries retrieve frequent itemsets, appearing in at least δ % of the records. The parameter δ varies in $\{0.625, 1.25, 2.5, 5\}$. To quantify the accuracy of answering the queries in a workload, we measured the average percentage of records that are retrieved incorrectly as part of the query answers on anonymized data. The latter percentage is denoted with α .

For example, consider a workload comprised of a query that returns 10 in the original data and 13 in the anonymized data. In this case, $(|10 - 13|/10 \cdot 100\% = 30\%$ of records are retrieved incorrectly as part of the query answer, thus $\alpha = 30$. Clearly, lower values of α imply higher data utility. Furthermore, we applied the CBA and DIS methods, using three different utility policies, namely \mathcal{U}_1 , \mathcal{U}_2 , and \mathcal{U}_3 . \mathcal{U}_1 is a *level 1 hierarchy-based* policy, \mathcal{U}_2 is a *similarity-based* policy, whose constraints contain 10 sibling 5-digit codes in the ICD hierarchy, and \mathcal{U}_3 is a *frequency-based* policy, whose constraints contain all codes that appear in at least 1.25 % of records. The parameters k and m were fixed to 5 and 2, respectively.

The results for the utility policy \mathcal{U}_1 are summarized in Table 7.13a. As one can observe, the scores for DIS and CBA are 15.5 and 5.5, on average, respectively. Thus, both algorithms perform reasonably well. However, DIS can generally allow enforcing k^m -anonymity with lower information loss. For instance, when $\delta = 5\%$, DIS permits approximately 16 times more accurate query answering than CBA. A similar observation can be made from Tables 7.13b and 7.13c, which illustrate the corresponding results for \mathcal{U}_2 and \mathcal{U}_3 , respectively. Of note, the difference of DIS and CBA with respect to α increases with the parameter δ . This suggests that DIS should be preferred over CBA when the queries correspond to relatively frequent combinations of diagnosis codes.

Table 7.13 Average percentage of records that are retrieved incorrectly, for workloads having different δ values and for: (a) \mathcal{U}_1 , (b) \mathcal{U}_2 , and (c) \mathcal{U}_3

$\delta(\%)$	α for CBA	α for DIS	$\delta(\%)$	α for CBA	α for DIS
0.625	16	13	0.625	43	13
1.25	15	6	1.25	36	6
2.5	20	2	2.5	29	2
5	10	0.6	5	1	0.6

(a)	(b)	
$\delta(\%)$	α for CBA	α for DIS
0.625	32	12
1.25	28	6
2.5	22	1
5	4	0.5

(c)

7.6 Future Directions

To increase the adoption of anonymization technologies, it is important to guarantee that medical studies can be performed accurately on anonymized data. The utility-constrained approach, surveyed in this chapter, is a first step towards achieving this objective, but there are various other directions awaiting to be explored. In the following, we discuss some of these directions. Specifically, in Sect. 7.6.1, we discuss alternative ways of formulating utility constraints, which are relevant to different uses of anonymized data. Next, in Sect. 7.6.2, we consider the issue of providing guarantees for the satisfaction of specified utility policies.

7.6.1 Different Forms of Utility Constraints

Our discussion so far focused on utility constraints that are modeled as sets of diagnosis codes. The intuition behind this formulation is that data transformation should not affect relations between diagnosis codes falling into the same utility constraint. However, there are alternative formulations that are worth investigating. In the following, we present three such formulations.

First, there are studies aiming to discover patterns of diagnosis codes, which have certain desired properties (e.g., a support that exceeds a certain threshold, or deviates from its expected value [3]). In many cases, publishing the patterns alone is insufficient (e.g., due to regulations that require patient-level data to be shared), or inappropriate (e.g., due to requirements of certain data recipients). Thus, it is important to formulate utility constraints, whose satisfaction will enable the discovery of patterns, and then anonymize the diagnosis codes with respect to such constraints. However, the problem is far from straightforward, due to the

overlap between patterns and the large number of patterns that may be discovered. Furthermore, generalization and/or disassociation must be applied in a way that prevents the creation of fake patterns (i.e., patterns that are not discoverable from the original dataset but can be discovered from the anonymized dataset).

Second, diagnosis codes are often released for querying purposes. In this case, the formulation of utility constraints should be based on the queries, in order to better preserve utility for the intended studies. This raises several interesting questions, including: (i) which types of queries can be supported best by generalization or disassociation, (ii) how to deal with correlations between diagnosis codes that must be retained to support the queries, and (iii) how to detect and deal with attackers who formulate queries in a way that makes breaching privacy easier. A pioneering attempt towards producing anonymized relational data (e.g., patient demographics) that support, but do not guarantee, accurate query answering was made in [8]. Diagnosis codes, however, require different treatment than demographics, and they are susceptible to different types of identity disclosure attacks [12].

Third, a large number of applications involve the analysis of both patient demographics *and* diagnosis codes. Methods for anonymizing such data have been considered recently [17, 18], but they do not guarantee the utility of anonymized data in intended studies. Formulating utility constraints for such data is worthwhile. However, it is not clear how to extend the utility-constrained approach to deal with data containing both demographics and diagnosis codes. For instance, numerical values contained in such data require different handling than diagnosis codes, as they are generalized using ranges, which are difficult to know prior to anonymization [26].

7.6.2 *Different Approaches to Guaranteeing Data Utility*

The utility-constrained approach treats privacy as more important than utility. That is, the utility policy may need to be violated, if this is necessary to satisfy privacy. In fact, no guarantees that a utility policy will be satisfied are provided by existing methods [11, 12, 15]. For instance, both CBA and DIS did not satisfy the utility policy, in the example of Sect. 7.5.3, although they satisfied some of the utility constraints (i.e., the constraints whose MRE is 0 in Table 7.12).

One way to get around this problem is to design algorithms that offer some type of guarantees, related to the satisfaction of utility constraints. Example of such guarantees are: “a predetermined subset of the utility constraints will always be satisfied”, or “the MRE of all utility constraints will be within a predetermined interval”. The challenge for the development of such algorithms stems from the computational complexity of the problem. As shown in [13], the problem of anonymizing data in the presence of utility constraints, formulated as sets of diagnosis codes, is NP-hard. An alternative way is to design more flexible privacy principles. Such principles may be based on the personal privacy preferences of patients, or employ different parameters (e.g., a lower k and m) for diagnosis codes, contained in utility constraints that are important to satisfy.

7.7 Conclusion

In this chapter, we provided a detailed discussion of the utility-constrained approach to anonymizing diagnosis codes. After introducing the operations of generalization and disassociation, we focused on the specification of utility constraints. Subsequently, we surveyed two algorithms that follow the utility-constrained approach. These algorithms are based on generalization and disassociation, respectively, and they have been shown to outperform others in terms of preserving data utility. Last, we discussed several promising directions for future work.

References

1. Denny, J.: Chapter 13: mining electronic health records in the genomics era. *PLoS Comput. Biol.* **8**(12), e1002823 (2012)
2. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: a survey on recent developments. *ACM Comput. Surv.* **42**(4), 1–53 (2010)
3. Gallo, A., De Bie, T., Cristianini, N.: Mini: mining informative non-redundant itemsets. In: *Principles and Practice of Knowledge Discovery in Databases (PKDD)*. Lecture Notes in Computer Science, pp. 438–445. Springer, Heidelberg (2007)
4. Gkoulalas-Divanis, A., Loukides, G.: In: *Anonymization of Electronic Medical Records to Support Clinical Analysis*. Springer Briefs in Electrical and Computer Engineering. Springer, New York (2013)
5. Gkoulalas-Divanis, A., Loukides, G., Sun, J.: Publishing data from electronic health records while preserving privacy: a survey of algorithms. *J. Biomed. Inform.* **50**(0), 4–19 (2014). Special Issue on Informatics Methods in Medical Privacy
6. He, Y., Naughton, J.F.: Anonymization of set-valued data via top-down, local generalization. *Proc. Very Large Data Bases Endowment* **2**(1), 934–945 (2009)
7. INFORMS Data Mining Contest (2008). <https://sites.google.com/site/informsdataminingcontest/>
8. LeFevre, K., DeWitt, D., Ramakrishnan, R.: Workload-aware anonymization. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 277–286 (2006)
9. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l-diversity. In: *International Conference on Data Engineering (ICDE)*, pp. 106–115 (2007)
10. Loukides, G., Denny, J., Malin, B.: The disclosure of diagnosis codes can breach research participants' privacy. *J. Am. Med. Inform. Assoc.* **17**, 322–327 (2010)
11. Loukides, G., Gkoulalas-Divanis, A.: Utility-aware anonymization of diagnosis codes. *IEEE J. Biomed. Health Informatics* **17**(1), 60–70 (2013)
12. Loukides, G., Gkoulalas-Divanis, A., Malin, B.: Anonymization of electronic medical records for validating genome-wide association studies. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 7898–7903 (2010)
13. Loukides, G., Gkoulalas-Divanis, A., Malin, B.: COAT: Constraint-based anonymization of transactions. *Knowl. Inf. Syst.* **28**(2), 251–282 (2011)
14. Loukides, G., Gkoulalas-Divanis, A., Shao, J.: Anonymizing transaction data to eliminate sensitive inferences. In: *Database and Expert Systems Applications (DEXA)*. Lecture Notes in Computer Science, pp. 400–415. Springer, Heidelberg (2010)
15. Loukides, G., Liagouris, J., Gkoulalas-Divanis, A., Terrovitis, M.: Disassociation for electronic health record privacy. *J. Biomed. Inform.* **50**(0), 46–61 (2014). Special Issue on Informatics Methods in Medical Privacy

16. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkatasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: IEEE International Conference on Data Engineering (ICDE), p. 24 (2006)
17. Mohammed, N., Jiang, X., Chen, R., Fung, B.C.M., Ohno-Machado, L.: Privacy-preserving heterogeneous health data sharing. *J. Am. Med. Inform. Assoc.* **20**(3), 462–469 (2013)
18. Poulis, G., Loukides, G., Gkoulalas-Divanis, A., Skiadopoulos, S.: Anonymizing data with relational and transaction attributes. In: The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery (ECML/PKDD), pp. 353–369 (2013)
19. Safran, C., Bloomrosen, M., Hammond, W., Labkoff, S., Markel-Fox, S., Tang, P., Detmer, D.: Toward a national framework for the secondary use of health data. *J. Am. Med. Inform. Assoc.* **14**, 1–9 (2007)
20. Samarati, P.: Protecting respondents identities in microdata release. *IEEE Trans. Knowl. Data Eng.* **13**(9), 1010–1027 (2001)
21. Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Uncertainty Fuzziness Knowledge Based Syst.* **10**, 557–570 (2002)
22. Terrovitis, M., Liagouris, J., Mamoulis, N., Skiadopoulos, S.: Privacy preservation by disassociation. *Proc. Very Large Data Bases Endowment* **5**(10), 944–955 (2012)
23. Terrovitis, M., Mamoulis, N., Kalnis, P.: Privacy-preserving anonymization of set-valued data. *Proc. Very Large Data Bases Endowment* **1**(1), 115–125 (2008)
24. Weiskopf, N.G., Weng, C.: Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J. Am. Med. Inform. Assoc.* **20**(1), 144–151 (2013)
25. Xiao, X., Tao, Y.: Anatomy: simple and effective privacy preservation. In: Proceedings of the 32nd International Conference on Very Large Data Bases, pp. 139–150 (2006)
26. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.W.C.: Utility-based anonymization using local recoding. In: In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data (KDD), pp. 785–790 (2006)
27. Xu, Y., Wang, K., Fu, A.W.C., Yu, P.S.: Anonymizing transaction databases for publication. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), pp. 767–775 (2008)

Chapter 8

Methods to Mitigate Risk of Composition Attack in Independent Data Publications

**Jiuyong Li, Sarowar A. Sattar, Muzammil M. Baig, Jixue Liu,
Raymond Heatherly, Qiang Tang, and Bradley Malin**

Abstract Data publication is a simple and cost-effective approach for data sharing across organizations. Data anonymization is a central technique in privacy preserving data publications. Many methods have been proposed to anonymize individual datasets and multiple datasets of the same data publisher. In real life, a dataset is rarely isolated and two datasets published by two organizations may contain the records of the same individuals. For example, patients might have visited two hospitals for follow-up or specialized treatment regarding a disease, and their records are independently anonymized and published. Although each published dataset poses a small privacy risk, the intersection of two datasets may severely compromise the privacy of the individuals. The attack using the intersection of datasets published by different organizations is called a composition attack. Some research work has been done to study methods for anonymizing data to prevent a composition attack for independent data releases where one data publisher has no knowledge of records of another data publisher. In this chapter, we discuss two exemplar methods, a randomization based and a generalization based approaches, to mitigate risks of composition attacks. In the randomization method, noise is added

J. Li (✉) • S.A. Sattar • J. Liu

School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, SA, Australia

e-mail: jiuyong.li@unisa.edu.au; Jixue.liu@unisa.edu.au; sarowar@gmail.com

M.M. Baig

InterSect Alliance International Pty Ltd, Adelaide Area, SA, Australia

e-mail: mirza-muzammil.baig@intersectalliance.com

R. Heatherly

Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

e-mail: r.heatherly@vanderbilt.edu

Q. Tang

APSIA group, SnT, University of Luxembourg, Walferdange, Luxembourg

e-mail: qiang.tang@uni.lu

B. Malin

Departments of Biomedical Informatics and EE and CS, Vanderbilt University, Nashville, TN, USA

e-mail: b.malin@vanderbilt.edu

to the original values to make it difficult for an adversary to pinpoint an individual’s record in a published dataset. In the generalization method, a group of records according to potentially identifiable attributes are generalized to the same so that individuals are indistinguishable. We discuss and experimentally demonstrate the strengths and weaknesses of both types of methods. We also present a mixed data publication framework where a small proportion of the records are managed and published centrally and other records are managed and published locally in different organizations to reduce the risk of the composition attack and improve the overall utility of the data.

8.1 Introduction

The increasing collection of large volumes of person information has created tremendous opportunities for knowledge-based decision making in a number of domains [29]. To fully maximize the knowledge that can be learned, the data needs to be made available beyond the organizations that performed the initial collection [20]. Data sharing, however, must be accomplished in a manner that respects the privacy of the individuals from which the data was gathered [35].

Data publication is an easy and effective way for data sharing. A wide variety of computational models have been proposed [13] for privacy preserving data publications. Most models focus on single publication [22, 23, 34] and multiple publications from the same data publisher [4, 14, 38]. Single publication refers to when a dataset is published once only and other possibly relevant datasets may have not been considered in the data publication process. Multiple data publications from the same data publisher refer to the publications of a series of datasets which are all extensions in some aspect (e.g., quarterly releases of data). When the datasets are all from the same publisher, the publisher has the knowledge of all original data.

In this chapter, we consider a situation where datasets are published independently from multiple organizations and they share records of some common individuals. This is a concern because, in some cases, an individual’s information will be collected and published by different organizations [24]. And when such a situation arises, an adversary may invoke a *composition attack* [15] on the published datasets to compromise the privacy guarantees of each dataset.

We illustrate the risk of a composition attack on multiple independent publications as follows. Suppose that a victim has the following personal information, (Gender = male, Age = 22, Postcode = 5095), which are known to an adversary. The adversary also knows that the victim’s records are in two datasets. Such information may be obtained through personal acquaintance or via social media. Table 8.1 lists data segments from the two published datasets containing the victim’s records. In each dataset, the adversary cannot uniquely determine the record corresponding to the victim’s sensitive information (i.e., disease in this case), because there are three possible diseases. However, the intersection of the two data segments contains only one disease, thus leading to the compromise of the victim’s

privacy. In this example, the sensitive value of the individual is 100 % inferred. Note that the confidence of the inference does not have to be 100 %, and, in practice, a highly confident inference can be considered a compromise of an individual's privacy.

Privacy risk in multiple independent data publications is real in practice. For example, a patient may visit two hospitals for the treatment of the same disease and the two hospitals then publish their datasets independently without coordination. The reasons for not coordinating may be due to regulations that disallow hospitals to share their raw data, or to infeasibility because a hospital may share common patients with many other hospitals [5].

There are a few methods dealing with composition attacks with multiple independent data publications. In this chapter, we discuss two methods for mitigating the risk of composition attacks at the conceptual level, analyze their strengths and weaknesses, and present a mixed publication framework for simultaneously reducing the risk of composition attacks and improving data utility.

The remainder of this chapter is organized as follows. Section 8.2 explains the composition attack and multiple publications in detail. Sections 8.3 and 8.4 discuss two methods for risk mitigation based on randomization and generalization, respectively. Section 8.5 presents experimental results to demonstrate the reduction of risk of composition attack by the two methods and compare their data utility. Section 8.6 provides a mixed publication framework for reducing the risk of the composition attack and achieving high data utility. Finally, Section 8.7 concludes the chapter.

8.2 Composition Attack and Multiple Data Publications

8.2.1 Composition Attack

A composition attack makes use of multiple anonymized datasets to determine the sensitive information of an individual whose records are contained in multiple datasets. The risk of composition attack is real in data publications, since datasets are rarely isolated. The problem is expected to become more severe when more datasets are publicly available.

An adversary is assumed to possess the following knowledge to make a composition attack possible. Firstly, the adversary knows *quasi-identifier attribute*

Table 8.1 An illustrative example of a composition attack; the shared common record is revealed by intersecting two corresponding equivalence classes

Gender	Age	Postcode	Disease
m	21–29	5091–5095	Anemia
m	21–29	5091–5095	Cancer
m	21–29	5091–5095	Diabetes

(a) An equivalence class in published dataset 1

Gender	Age	Postcode	Disease
m	21–29	5091–5099	Cancer
m	21–29	5091–5099	Migraine
m	21–29	5091–5099	Cough

(b) An equivalence class in published dataset 2

values of an individual (i.e., victim) whose sensitive value is to be inferred. Quasi-identifier attributes are attributes that, in combination, can potentially disclose the identity of an individual, such as `age`, `sex`, and `zip` code of residence. Secondly, the adversary knows that the record of a victim is stored in two or more datasets that are anonymized and published.

The mechanism for a composition attack is an intersection operation. In an anonymized dataset, records are normally organized into small groups, each with the identical set of quasi-identifier values, and such a group is called an *equivalence class*. In an equivalence class, all individuals are indistinguishable and they are associated with some number of sensitive values, dependent upon the specific method of anonymization. For example, Table 8.1(a) contains an equivalence class which is associated with three sensitive values ‘Anemia’, ‘Cancer’, ‘Diabetes’. As a result, an adversary is unsure as to the sensitive value of a victim, even if she can identify the equivalence class that the victim is in. Hence, the privacy of a victim is protected.

However, if the adversary knows that two equivalence classes in two datasets contain the victim’s record, she can use the intersection to narrow down the set of sensitive values in common in both equivalence classes. If there is only one sensitive value in common in the two equivalence classes, the privacy of the victim is completely compromised. Table 8.1 shows such a case, since only one sensitive value ‘Cancer’ is common in both equivalence classes. When there is only one common sensitive value, the confidence of a composition attack is 100 %. More generally, the privacy of an individual can be considered to be compromised if adversary’s confidence is substantially larger than a random guess. When an individual’s record is mapped to more than one equivalence class, the set of equivalence classes form a super equivalence class. Yet, the privacy can be compromised by applying the mechanism of the composition attack to the super equivalence classes. We consider the foundational case where one record maps to one equivalence class in this chapter.

The aim for protecting datasets from composition attacks is to ensure that the common sensitive values in two or more overlapping equivalence classes are at least d , where d is the user-specified minimum number of sensitive values and $d \geq 2$. We say that two equivalence classes *overlap* if their corresponding values are identical or their value ranges overlap. For example, two equivalence classes in Table 8.1(a) and (b) are overlapping since they have the same `Gender` value and `Age` intervals, and their `Postcode` ranges overlap respectively.

Before we move to multiple data publications, we briefly recap two most widely used anonymization models in single data publication. k -anonymity [34] requires the minimum size of an equivalence class in a published dataset at least k so an adversary has at most $1/k$ confidence in pinpointing the record of a victim. l -diversity [23] requires the number of sensitive values in an equivalence class to be at least l and, hence, the confidence of an adversary inferring the sensitive value of a victim be $1/l$ (assuming that the l values have the same frequencies). For example, both tables in Table 8.1 are 3 anonymous and 3 diverse.

8.2.2 *Multiple Coordinated Data Publications*

In multiple coordinated data publications, a data publisher knows the raw and published datasets of all versions, or accesses the published datasets of other organizations in the data anonymization process. Serial data publication [14, 36, 38, 39] is a typical example. For example, a hospital publishes quarterly patient activity reports, and it makes use of all information in the previous versions when it anonymizes its current version of the dataset.

Serial publications have attracted some research interest in the privacy-preserving data publications [14, 36, 38, 39]. A widely accepted model is *m*-invariance [39], which requires all overlapping equivalence classes in all versions of published datasets persistently associate with the same set of *m* sensitive values. As a result, when an adversary tries to use an intersection to narrow down the number of sensitive values using two or more equivalence classes in different versions, the minimum number of common sensitive values she can get is *m*, and hence the privacy of individuals is protected from composition attacks. Note that the previously published datasets are taken into account when anonymizing the current dataset.

Collaborative privacy-preserving data publication solves a different problem. It concerns two or more parties, each holding a part of data, who wish to collaboratively publish an integrated and anonymized dataset [13]. The techniques in this area generally fall in two categories: *collaborative anonymization for vertically partitioned data* [16, 25] and *collaborative anonymization for horizontally partitioned data* [17, 41]. In the former category, different data publishers hold datasets of different attributes while the records may or may not correspond to the same entity. In the latter category, different data publishers hold different entities, each with the same set of attributes. In collaborative privacy-preserving data publications, it is essential to access the datasets of other parties in a secure manner (e.g., via secure multi-party communication).

The methods for collaborative privacy-preserving data publications can, however, be used for multiple coordinated data publications, especially different datasets belong to different publishers who do not share their raw data. Following a privacy-preserving serial data publication model, such as the *m*-invariance model, different data publishers can use secure computation or build joint generalization taxonomies to ensure that in the anonymization process the principle is satisfied.

8.2.3 *Multiple Independent Data Publications*

In multiple independent data publications, a data publisher does not know other datasets (published or to be published) that might be used for composition attacks. The data publisher does not access other datasets in the data anonymization process.

Simply speaking, in the process of an independent data publication, a data publisher does not reference any other published or unpublished dataset.

It is very difficult to protect the privacy in a dataset from composition attacks in multiple independent data publications, since datasets to be used for composition attacks are unknown. There is little work on data anonymization for multiple independent data publication.

The first work studying the privacy risk on multiple independent data publications was reported in [15]. In addition to demonstrate the privacy risks of multiple independent data publications, the authors prove that the concept of differential privacy can be used to protect data from composition attacks. In [15], it is unclear whether the differential privacy based solution is for data publications or for interactive data access (where a user makes queries to an unpublished database and receives perturbed results). Nevertheless, the conclusion in [15] is true for differential privacy based data publications. A concern with this solution is that the offered data utility is low. Later on in the chapter, we will provide a detailed discussion on the privacy and utility of differential privacy based data publications.

Another method to anonymize datasets to reduce the risk of composition attacks in multiple independent publications is by *cloning* [2]. The method works under the assumption that all datasets are drawn from the same population and their distributions are similar. In an idea situation, assume that the distributions of sensitive values in all datasets are identical, and a dataset is anonymized in a way that each equivalence class has the identical sensitive value distribution as the dataset. In this idea situation, the risk of composition attack is zero. As an approximation, if the frequency difference of each sensitive value between an equivalence class and the dataset is bounded by a small value, the risk of composition attack will also be bounded. The work by Baig et al. [2] implements an approximate algorithm for cloning generalization. The limitation of this method is that the size of an equivalence class can be large and this reduces the data utility.

A more recent method to solving the problem is to use a probabilistic approach to control the risk of composition attack [32]. Based on the assumption that all datasets are drawn from the same population and their distributions are similar, the chance of a sensitive value falling into an equivalence class can be estimated and the chance of two equivalence classes sharing a number of sensitive values can be estimated as well. The assumption is realistic in real word applications, because the disease distribution and demographic distributions of patients are largely similar. If the chance of sharing a minimum number of sensitive values in two or more equivalence classes in different datasets is greater than a user specified threshold, then the chance of a composition attack is bounded. The limitation of the model is that low frequency sensitive values may result in significant utility loss, since records with them need a high level generalization.

In the following, we further explain the first and third methods as examples for randomization and generalization approaches for reducing the risk of composition attacks in multiple independent publications, and elaborate their strengths and weaknesses. We then present a mixed publication framework for data publications to mitigate the risk of composition attacks.

8.3 Risk Mitigation Through Randomization

Randomization is a major technique for data anonymization. A randomization method adds noise to the original values (or aggregated values), and makes it difficult to pinpoint an individual in a published dataset or to infer the exact sensitive value of an individual. In general, it creates uncertainty in a published dataset and reduces the probability of inferring the sensitive information of an individual.

Differential privacy is a major framework in randomization. A randomized mechanism on a dataset satisfies differential privacy if the removal or inclusion of a single record from the dataset has only a small effect on the output of the randomization mechanism [8]. It was proposed in an interactive setting, where a user accesses a database via a query and an anonymization technique adds noise to the query result. To ensure privacy, the user is given a privacy budget and is allowed to query the database until the budget is exhausted. However, the interactive setting is not always applicable since, in some situations, datasets are required to be published publicly to provide more flexible analysis [26].

The differential privacy principle has been applied to non-interactive data access. A data set can be summarized as a contingency (or frequency) table, where each row takes a set of unique values of attributes representing an equivalence class and the sensitive values are summarized as counts. A count query can be answered by aggregating relevant rows in the contingency table. The noise is added to the counts of the contingency table and the contingency table (or aggregations to some attributes) is published [3, 11]. One problem is that the number of rows of the contingency table can be large and many counts are small or zero, and noise on the small counts greatly reduces the data quality. The method in [26] generalizes attribute values and reduces the problem with small counts. In other words, a dataset published by Mohammed et al. [26] is row-wise aggregation of the original contingency table and counts of sensitive values are added Laplacian noise. However, the problem of small counts still exists, since the attribute values should not be generalized too much to preserve good utility of the published datasets for answering count queries. In this chapter, we consider the differential privacy based data publications, where values have been generalized.

A differential privacy based randomization method can protect data privacy from composition attacks [15]. When two datasets are processed with a randomized method of differential privacy, with privacy budgets ϵ_1 and ϵ_2 (the smaller a privacy budget, the higher privacy), a composition attack could not completely compromise the privacy of an individual, since the privacy is ensured in the privacy budget $\epsilon_1 + \epsilon_2$ for intersections of two datasets. An intuitive illustration of this is shown in Table 8.2 by contrast with Table 8.1. In Table 8.1 we have two independent publications using the standard anonymization techniques of k -anonymity [34] and l -diversity [23]. The pair of 3-anonymous and 3-diverse datasets are vulnerable to a composition attack since only one sensitive value, ‘Cancer’, is common between the two equivalence classes. The record of an individual will then be revealed if the two records correspond to the same individual. Table 8.2 shows differential privacy

based publications [15] of the same datasets. The pair of differential privacy based equivalence classes are not vulnerable to a composition attack, since an adversary cannot infer which counts are actually zero and which are not. A zero count may not be zero since a non-zero value may be noised to zero, and a non-zero count may be noised from a zero count. The intersection of sensitive values of two equivalence classes may not give the adversary the true common sensitive values.

The utility of a differential privacy based dataset can be low. The level of noise of a differential privacy based randomized method is independent from the value to be noised and is dependent on the user specified privacy budget ϵ . In other words, given a privacy budget ϵ , a small value and a large value in a table have the same chance to receive the same magnitude of noise. Relatively speaking, the impact of the noise is larger for a small value than for a large value. This is why differential privacy does not provide good utility for small query results. Other discussions on the low utility of differential privacy can be found in [7, 30, 33]. In differential privacy based publications [26], records are grouped into equivalence classes. The size of an equivalence class is not large because the detailed information of the quasi-identifier would otherwise be lost. So, the counts of sensitive values in an equivalence class are not large, and this results in a big impact of noise on the published dataset. For example, Table 8.2 shows that the counts of sensitive values do not preserve the original distributions of sensitive values in two equivalence classes well.

It is possible to increase ϵ to reduce the level of noise to maintain the utility, but this makes the datasets vulnerable to a composition attack. In a composition attack, the privacy is protected within the budget $\epsilon_1 + \epsilon_2$. If both ϵ_1 and ϵ_2 are large, the total budget may not satisfy the user's privacy requirement. For example, if the noise is low in Table 8.2, most of the zero counts will be published unaltered, and so the adversary is again able to infer with a high confidence that 'Cancer' is the only common sensitive value. Experiments reported in [32] show that when the value of ϵ is larger than 1, more than 75 % of the nosed values are unchanged (note that fractional values are rounded to their nearest integers).

In addition, noise may cause some other problems in medical practice. For example, when the count of a sensitive value within an equivalence is small in

Table 8.2 An illustration of differential privacy based publications of datasets in Table 8.1. Anemia, Cancer, Migraine, Diabetes and Cough are all sensitive values in the datasets. The counts of sensitive values are noised and published with the equivalence class. It is difficult for an adversary to find true common sensitive values using noised counts. Note that the counts are small since we use the same datasets from Table 8.1

G	Age	Postcode	Disease(counts)
m	21–9	5091–99	Anemia (2)
m	21–9	5091–99	Cancer (1)
m	21–9	5091–99	Migraine (1)
m	21–9	5091–99	Diabetes (1)
m	21–9	5091–99	Cough (0)

(a) An equivalence class in published dataset 1 (G for Gender)

G	Age	Postcode	Disease(counts)
m	21–9	5091–99	Anemia (0)
m	21–9	5091–99	Cancer (1)
m	21–9	5091–99	Migraine (0)
m	21–9	5091–99	Diabetes (1)
m	21–9	5091–99	Cough (1)

(b) An equivalence class in published dataset 2 (G for Gender)

the original dataset, say zero, it may be perturbed into any integer. And this is problematic when analytics over the published data require the complete absence of information in order to classify an individual, such as is the case in clinical phenotyping [28]. Even more, this problem can occur for all equivalence classes of individuals. Overall, the chance for a result to have a large noise is low. However, for a user making just one query, the worst case is that the result has been added a large amount of noise and hence is useless. So, many practitioners prefer generalization to randomization, since the uncertainty in the generalization is bounded for a result but the uncertainty of a noised result is unbounded for a single result (the noise can be anything between zero and a large value).

8.4 Risk Mitigation Through Generalization

Generalization is another main technique for data anonymization. Generalization coarsens the values in the quasi-identifier attributes and makes a number of individuals appear to be identical in a published dataset. As a result, an individual cannot be identified and her sensitive value cannot be inferred with a high confidence. Generalization has been used for the implementation of most partition-based privacy-preservation models, such as k -anonymity [34], l -diversity [23], t -closeness [22], and (α, k) -anonymity [37]. It has also been used to implement anonymization models for serial publications [14, 36, 38, 39]. Note that the number of quasi-identifier attributes should not be large for generalization to be effective, because as the number of attributes grows, so too does the amount of generalization in order to achieve privacy protection, which leads to low data utility.

In general, generalization methods do not protect data privacy from composition attacks in multiple independent publications [1]. In a generalized dataset, the information is coarsened but faithful. When an adversary notices that the victim's quasi-identifier values match the generalized values of an equivalence class, the victim's record is in the equivalence class. The counts of sensitive values in an equivalence class are also faithful too. When a count is zero, the sensitive value does not occur in the equivalence class. Since two datasets are published independently, there is always a chance (even though very small) that two equivalence classes share only one sensitive value and hence a composition attack is possibly successful (when the matched records belong to the victim whom the adversary is searching for).

To address the composition attacks in multiple independent data publications via generalization, we will have to consider a probabilistic model, which allows a small chance of possible composition attacks. Even in differential privacy based publications, there is a chance for possible composition attacks. For a stubborn adversary who takes everything she sees in differentially private tables as true values, she will find unique common sensitive value and this finding has a chance to be true. The protection of data from composition attacks by differential privacy does not mean there is no possibility for a successful composition attack, but the chance is low.

Table 8.3 Comparison of risk of composition attack of two equivalence classes: Case 1

G	Age	Postcode	Disease
m	20–9	5090–99	Leukaemia
m	20–9	5090–99	Huntington's disease
m	20–9	5090–99	Multiple sclerosis

(a) Equivalence class 1 (G for Gender)

G	Age	Postcode	Disease
m	20–9	5090–99	Cold
m	20–9	5090–99	Migraine
m	20–9	5090–99	Diabetes

(b) Equivalence class 2 (G for Gender)

Since other datasets are not referenced in multiple independent data publications, it is impossible to use a serial or collaborative method to ensure the minimum number of common sensitive values in all possible intersections of corresponding equivalence classes in two published datasets. The best we can do is to ensure that in most intersections, the minimum number of common sensitive values are maintained. To motivate our discussion, let us look at two equivalence classes in Table 8.3. Without other information, which equivalence class is more risky for a composition attack? Note that we do not know the table to be used for a composition attack.

Intuitively, the equivalence class in Table 8.3(a) has a higher risk for a composition attack than the equivalence class in Table 8.3(b). The chance of any pair of sensitive values in Table 8.3(b) occurring in an equivalence class of another table is higher than that of any pair in Table 8.3(a). This is because ‘Cold’, ‘Migraine’ and ‘Diabetes’ are common diseases and the chance to see any pair of them in an equivalence class of a middle age group is high. All three diseases in Table 8.3(a) are uncommon and seeing any two of them in one equivalence class is rare. In other words, the sensitive values in Table 8.3(a) have a higher chance of being unique when the equivalence class is intersected with another equivalence class, than those in Table 8.3(b). So the risk of the equivalence class in Table 8.3(a) is higher than that of the equivalence class in Table 8.3(b).

The risk for a composition attack is associated with the frequencies of sensitive values in a dataset. The higher frequencies of sensitive values, the lower the chance for a composition attack.

Again, let us look at the two equivalence classes shown in Table 8.4. All diseases in Table 8.4(a) and (b) are uncommon. However, Table 8.4(b) has a lower chance for a composition attack than Table 8.4(a), since there are more sensitive values in Table 8.4(b) than in Table 8.4(a). The chance of another equivalence class containing two sensitive values of Table 8.4(b) is higher than that of Table 8.4(a), and this reduces the chance of a possible composition attack.

The risk for a composition attack is associated with the number of sensitive values in an equivalence class. The more sensitive values, the lower the chance for a composition attack. Therefore, without the knowledge of a dataset to be used for a composition attack, we are still able to reduce the probability of the attack, as long as *each equivalence class contains a sufficient number of sensitive values according to their frequencies*.

The (d, α) -linkable model [32] was designed to mitigate the risk of composition attack in multiple independent publications. We provide an illustrative explanation

Table 8.4 Comparison of risk of composition attack of two equivalence classes: Case 2

G	Age	Postcode	Disease
m	20–9	5090–99	Leukaemia
m	20–9	5090–99	Huntington’s disease
m	20–9	5090–99	Multiple sclerosis
(a) Equivalence class 1 (G for Gender)			
G	Age	Postcode	Disease
m	20–9	5090–99	Leukaemia
m	20–9	5090–99	Huntington’s disease
m	20–9	5090–99	Multiple sclerosis
m	20–9	5090–99	Lung cancer
m	20–9	5090–99	Prostate cancer
(b) Equivalence class 2 (G for Gender)			

of the (d, α) -linkable model in the following. Assume that all datasets are drawn from a population and they have the same distributions in quasi-identifier attributes and in the sensitive attribute. It is possible to estimate the probability of sharing d (i.e., the user-specified minimum number of common sensitive values of two overlapping equivalence classes) of any two datasets drawn from the population, and this probability is denoted as α_1 . Then, $1 - \alpha_1$ is the probability of two overlapping equivalence classes sharing sensitive values less than d , and it poses a risk for composition attacks when $(1 - \alpha_1) > (1 - \alpha)$, where α is a user-specified minimum confidence threshold. When $\alpha_1 > \alpha$, the risk of composition attack is lower than the user-specified confidence level and hence the privacy is protected. Note that the shared sensitive values are not all directly caused by common records of same individuals; the majority of them are caused by chance (two irrelevant individuals happen to have the same generalized quasi-identifier attribute values and disease), and hence the probability for a possible composition attack is multiple magnitude lower than $(1 - \alpha_1)$.

Sattar et al. [32] also present a dLink algorithm as a generalization implementation of the (d, α) -linkable model. The dLink algorithm works as a post-processing procedure for a dataset, which is anonymized by a normal generalization algorithm protecting privacy in a single data release, in order to reduce the risk of composition attack. The experimental results show that the dLink process reduces the risk of composition attacks greatly and largely preserves data utility.

A major limitation of the dLink process is that it cannot handle well data with a skewed sensitive value distribution. For example, if two or more rare sensitive values are included in one equivalence class, the size of equivalence class has to be very large to satisfy the (d, α) -requirement, and this results in significant utility loss. We will consider this issue in the final section of this chapter.

8.5 An Experimental Comparison

In this section, we demonstrate the reduction of risk of composition attacks by the randomization based approach and the generalization based approach discussed in the previous section. We compare the utility of these methods in terms of query accuracy and preservation of the distribution of sensitive values.

Table 8.5 Domain size of different attributes

Attribute	Age	Sex	Education	Race	Birth place	Occupation	Salary
Domain Size	100	2	20	6	41	50	50

8.5.1 Data and Setting

We performed experiments with real world datasets derived from the U.S. Census Bureau.¹ We split the dataset into two independent datasets (a) *Occupation* and (b) *Salary*. Each dataset consists of 600,000 records. The *Occupation* dataset includes five quasi-identifier attributes: Age, Gender, Education, Race, Birth-place and one sensitive attribute: Occupation. The *Salary* dataset contains the same quasi-identifier attributes and the sensitive attribute: Salary. All quasi-identifier attributes consist of categorical values except Age and Education. A description of the datasets is shown in Table 8.5.

We created five sets of datasets with overlapping records in the following way. We firstly initiated five disjoint data groups from every dataset, each with 100,000 records. The remaining 100,000 records are used for an overlapping record pool. We then made five copies of each group, and purposely inserted 1000, 2000, 3000, 4000 and 5000 records from the overlapping pool to the copies respectively for each group. We finally obtained five sets of datasets with the size of 101,000, 102,000, 103,000, 104,000 and 105,000. Each set of datasets shares overlapping records of 1000, 2000, 3000, 4000 and 5000, respectively.

The randomization based data publication method used in the experiment is the implementation of the differential privacy framework proposed in [26]. Privacy budgets ϵ are set as 0.05 and 0.1. In the experiments of most other works, ϵ is set between 0.1 and 0.2. Consider that the privacy budget after the composition attack using two datasets is an aggregation of two individual privacy budgets. We use 0.05 and 0.1. To give readers an idea of the noise level of $\epsilon = 0.05$ and 0.1, they leave around 20 and 30 % values un-noised [31].

We apply the post-processing method (i.e., dLink) to datasets that have been k -anonymized by the Mondrian algorithm [21]. The value of k varies from 10 to 30.

8.5.2 Reduction of Risk of Composition Attacks

Composition attacks are conducted between all pairs of datasets with the same overlapping records. The risk is measured as the accuracy of composition attack, which is defined as the ratio of the number of records have one common sensitive

¹<http://ipums.org>.

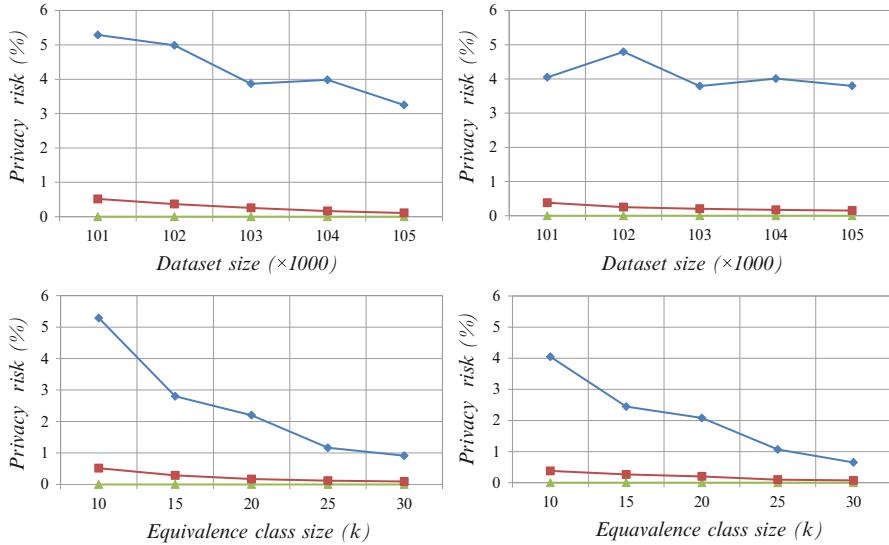


Fig. 8.1 The average accuracy of the composition attack on the *Salary* and *Occupation* datasets

value in its corresponding equivalence classes of different anonymous datasets and total number of overlapping records. The results are averages of all pairs of datasets.

Figure 8.1 shows privacy risk of the composition attack of both the differential privacy based method and the generalization based method by post-processor dLink in comparison with publications by a k -anonymization method (Mondrian [21]). Both methods have reduced the risk of composition attack greatly. The risk of composition attack of the differential privacy based method is nearly zero. Consider that the number of sensitive values is 50 and the count of the majority of them are noised. The chance of 1 common sensitive value is very low. dLink reduces the success rate of the composition attack greatly. The risk reduction increases with the size of the equivalence class. In some cases, the privacy risk is lower than one-tenth of the privacy risk without the post-processor (dLink). The risk of composition attack is nonzero especially when k is small. This is because the number of sensitive values in each equivalence class is even smaller which results in a nonzero probability of the unique common sensitive value in two overlapping equivalence classes. Moreover, from Fig. 8.1 it can be observed that the privacy risk of Mondrian and dLink approaches zero when the equivalence class size is large. The reason behind this effect is that when two equivalence classes are large, the likelihood that there are common sensitive values in two equivalence classes increases significantly. As a consequence, the risk of composition attack is reduced. However, this does not mean that it is unnecessary to utilize dLink. First, if we increase equivalence class size to reduce the risk of the composition attack, we unnecessarily reduce the utility of dataset. Second, a large equivalence class does not directly bound the risk of composition attack.

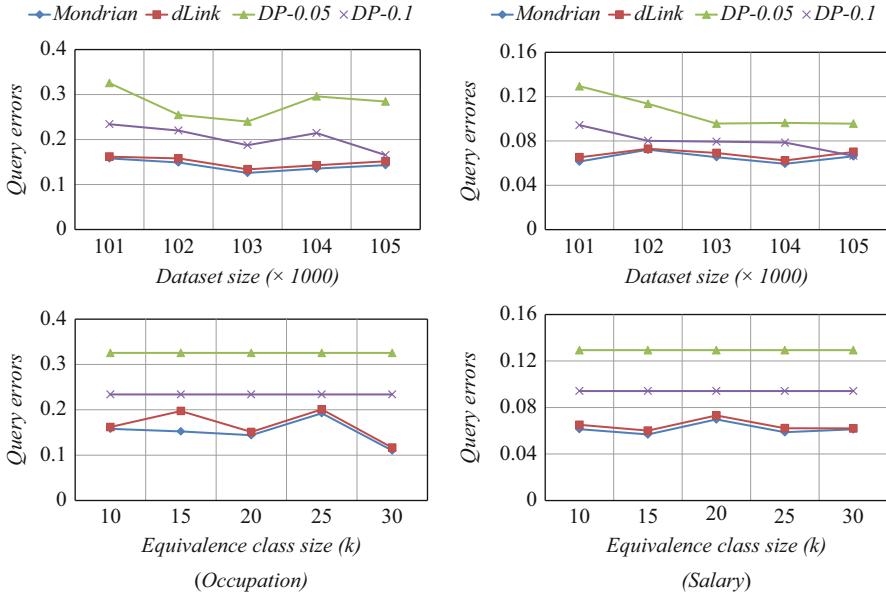


Fig. 8.2 The average query errors of the *Salary* and *Occupation* datasets with different methods

8.5.3 Comparison of Utility of the Two Methods

We assess the utility of published datasets by the accuracy of answering range queries. We randomly generate 1000 queries. For a query, we obtain its true result from the original dataset, and compute an estimated answer from its anonymized dataset. The relative error of a query is defined as $RE = (|\text{True result} - \text{Estimated result}|)/(\text{True result})$. We measure the workload error as the average relative error (ARE) of all the queries of all datasets.

Figure 8.2 lists the average query errors of the k -anonymized datasets with and without dLink, and of differential privacy anonymized datasets. The results show that relative difference in query error is at most 5 % before and after the application of dLink. It can be observed that in all cases, queries on dLink processed datasets have at least 7 and 22 % less error than those from differential privacy anonymized datasets with ϵ of 0.1 and 0.05, respectively. Note that the query errors do not change with k for differential privacy based anonymization, since k is not a parameter in the implementation. The data set with different k is of size of 101,000.

The above results support our previous analysis that *the differential privacy based approach does not preserve data utility well*. In contrast, the post-processing strategy that is applied by dLink has little effect on query accuracy. It reduces the data utility slightly (5 % or less errors in the query processing) in comparison with k -anonymized datasets.

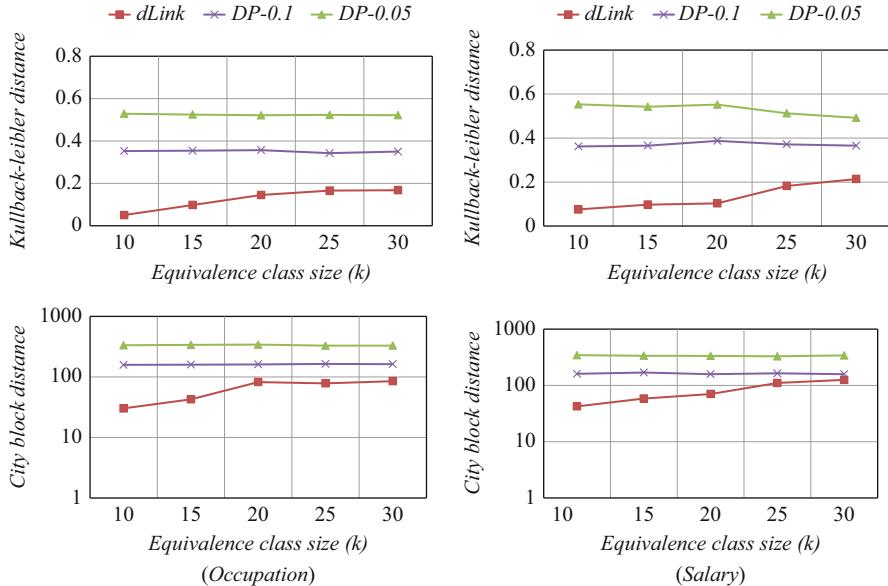


Fig. 8.3 Distance between the original dataset, the output of dLink, and several privacy budgets of differential privacy ($\epsilon = 0.01, 0.05, 0.1$)

We also compare the distances of distributions (histograms) of sensitive values in equivalence classes in the published dataset and the original dataset without the anonymization. The distance is measured by the Kullback-Leibler distance [19] and city block distance. The smaller the distance, the better the preservation of the original distribution. The results in Fig. 8.3 are average distances of equivalence classes. From Fig. 8.3, it can be seen that the distributional distances using dLink are at least 21.03, and 56.51 % smaller than those using differential privacy based anonymization, when ϵ is set to 0.1 and 0.05, respectively.

8.6 Risk Mitigation Through Mixed Publications

In this section, we propose a mixed publication framework, combining independent publication and centralized publication schemes, to reduce risk of composition attack and improve data utility.

We prefer generalization to randomization in data publications, since generalization gives faithful information even though the granularity may not be fine, whereas the noised information may bother many users. Most information consumers do not need detailed information, but instead use aggregated information. For instance, medical researchers can query which age group “20–30” or “30–40” is more susceptible to Diabetes instead of an exact age value. Binning age values intervals

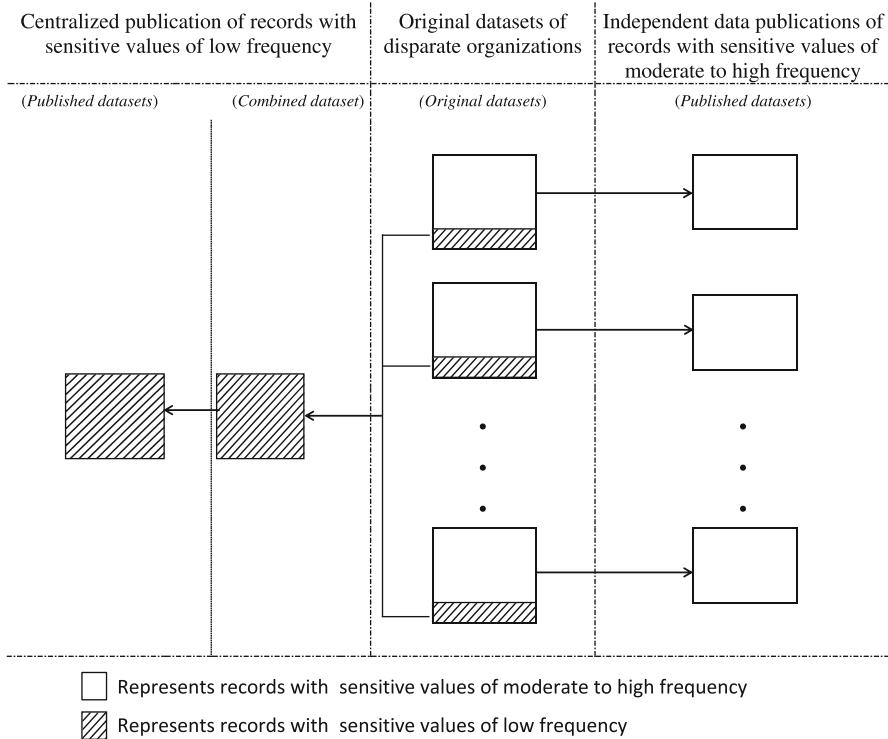


Fig. 8.4 An illustration of the mixed publication model

is a common practice in most applications [6]. So generalization does not lose much utility for many users. Though on average the noise added to a randomized dataset is small by a randomized method, if a user only takes one query result, the user does not know the impact of noise since noise may be anything between zero and a large amount. It is impossible for a user to make a number of same queries and to obtain the average query result in a data publication scenario.

The generalization based on the (d, α) -linkable model has been shown to greatly reduce the risk of composition attacks in the previous subsection and to preserve data utility well. However, the inherent limitation is that a skewed dataset may result in a substantial utility loss. We propose a mixed publication model to further reduce risk and improve utility in data publications.

The low frequency sensitive values are a major obstacle for processing data to reduce the risk of composition attack and maintain good data utility. If we try to accommodate all low frequency sensitive values in the (d, α) -model to keep the probability of d common sensitive values occurring in two equivalence classes high, then we will need a very high level of generalization to obtain large size equivalence classes. This sacrifices the utility of published datasets.

Centralized publication of records with sensitive values of low frequency	Original datasets of disparate organizations	Independent data publications of records with sensitive values of moderate to high frequency																																																																																																												
	<table border="1"> <thead> <tr> <th>Gender</th><th>Age</th><th>Postcode</th><th>Disease</th></tr> </thead> <tbody> <tr><td>m</td><td>25</td><td>5095</td><td>Flu</td></tr> <tr><td>m</td><td>31</td><td>5076</td><td>Migraine</td></tr> <tr><td>m</td><td>38</td><td>5055</td><td>Diabetes</td></tr> <tr><td>m</td><td>28</td><td>5026</td><td>Leukaemia</td></tr> </tbody> </table> <table border="1"> <thead> <tr> <th>Gender</th><th>Age</th><th>Postcode</th><th>Disease</th></tr> </thead> <tbody> <tr><td>m</td><td>26</td><td>5085</td><td>Cough</td></tr> <tr><td>m</td><td>35</td><td>5126</td><td>Migraine</td></tr> <tr><td>m</td><td>34</td><td>5150</td><td>Diabetes</td></tr> <tr><td>m</td><td>47</td><td>5066</td><td>HIV</td></tr> </tbody> </table> <table border="1"> <thead> <tr> <th>Gender</th><th>Age</th><th>Postcode</th><th>Disease</th></tr> </thead> <tbody> <tr><td>m</td><td>24</td><td>5295</td><td>Migraine</td></tr> <tr><td>m</td><td>38</td><td>5203</td><td>Cough</td></tr> <tr><td>m</td><td>33</td><td>5267</td><td>Flu</td></tr> <tr><td>m</td><td>50</td><td>5211</td><td>Amylose</td></tr> </tbody> </table>	Gender	Age	Postcode	Disease	m	25	5095	Flu	m	31	5076	Migraine	m	38	5055	Diabetes	m	28	5026	Leukaemia	Gender	Age	Postcode	Disease	m	26	5085	Cough	m	35	5126	Migraine	m	34	5150	Diabetes	m	47	5066	HIV	Gender	Age	Postcode	Disease	m	24	5295	Migraine	m	38	5203	Cough	m	33	5267	Flu	m	50	5211	Amylose	<table border="1"> <thead> <tr> <th>Gender</th><th>Age</th><th>Postcode</th><th>Disease</th></tr> </thead> <tbody> <tr><td>m</td><td>20-40</td><td>5000-5100</td><td>Flu</td></tr> <tr><td>m</td><td>20-40</td><td>5000-5100</td><td>Migraine</td></tr> <tr><td>m</td><td>20-40</td><td>5000-5100</td><td>Diabetes</td></tr> </tbody> </table> <table border="1"> <thead> <tr> <th>Gender</th><th>Age</th><th>Postcode</th><th>Disease</th></tr> </thead> <tbody> <tr><td>m</td><td>25-45</td><td>5000-5200</td><td>Cough</td></tr> <tr><td>m</td><td>25-45</td><td>5000-5200</td><td>Migraine</td></tr> <tr><td>m</td><td>25-45</td><td>5000-5200</td><td>Diabetes</td></tr> </tbody> </table> <table border="1"> <thead> <tr> <th>Gender</th><th>Age</th><th>Postcode</th><th>Disease</th></tr> </thead> <tbody> <tr><td>m</td><td>20-40</td><td>5200-5300</td><td>Migraine</td></tr> <tr><td>m</td><td>20-40</td><td>5200-5300</td><td>Cough</td></tr> <tr><td>m</td><td>20-40</td><td>5200-5300</td><td>Flu</td></tr> </tbody> </table>	Gender	Age	Postcode	Disease	m	20-40	5000-5100	Flu	m	20-40	5000-5100	Migraine	m	20-40	5000-5100	Diabetes	Gender	Age	Postcode	Disease	m	25-45	5000-5200	Cough	m	25-45	5000-5200	Migraine	m	25-45	5000-5200	Diabetes	Gender	Age	Postcode	Disease	m	20-40	5200-5300	Migraine	m	20-40	5200-5300	Cough	m	20-40	5200-5300	Flu
Gender	Age	Postcode	Disease																																																																																																											
m	25	5095	Flu																																																																																																											
m	31	5076	Migraine																																																																																																											
m	38	5055	Diabetes																																																																																																											
m	28	5026	Leukaemia																																																																																																											
Gender	Age	Postcode	Disease																																																																																																											
m	26	5085	Cough																																																																																																											
m	35	5126	Migraine																																																																																																											
m	34	5150	Diabetes																																																																																																											
m	47	5066	HIV																																																																																																											
Gender	Age	Postcode	Disease																																																																																																											
m	24	5295	Migraine																																																																																																											
m	38	5203	Cough																																																																																																											
m	33	5267	Flu																																																																																																											
m	50	5211	Amylose																																																																																																											
Gender	Age	Postcode	Disease																																																																																																											
m	20-40	5000-5100	Flu																																																																																																											
m	20-40	5000-5100	Migraine																																																																																																											
m	20-40	5000-5100	Diabetes																																																																																																											
Gender	Age	Postcode	Disease																																																																																																											
m	25-45	5000-5200	Cough																																																																																																											
m	25-45	5000-5200	Migraine																																																																																																											
m	25-45	5000-5200	Diabetes																																																																																																											
Gender	Age	Postcode	Disease																																																																																																											
m	20-40	5200-5300	Migraine																																																																																																											
m	20-40	5200-5300	Cough																																																																																																											
m	20-40	5200-5300	Flu																																																																																																											

Fig. 8.5 An example of the mixed publication

As such, the low frequency sensitive values and medium-to-high frequency sensitive values should be dealt with separately in data publications. We propose a mixed data publication model for reducing the privacy risk from composition attacks. For sensitive values with moderate to high frequency, they can be published independently by individual data publishers. For low frequency sensitive values, they should be aggregated further among a number of (or all) data publishers. In practice, there are such mechanisms already, for example, in national cancer registries. An illustrative diagram of the model is shown in Fig. 8.4. In addition, Fig. 8.5 shows an example of data publication following the model from three different organizations.

There are a number of benefits for such a mixed model:

1. It balances the cost for data publications, privacy protection and data utility. If all patient records are centralized, there will be too much data and the centralization is too costly to maintain. If all data records are locally managed and published, the low frequency diseases will cause problems for privacy or utility. The centralized management of rare diseases is not too costly, and makes the data of rare diseases precise and easy to analyze. Since the rare diseases in a local dataset are very few, it does not have much value for the analysis at the local level. Furthermore, if records of rare diseases are locally managed, it is unavoidable to have duplicated records since a patient may visit a number of hospitals. The small number of duplicates will have a big impact in data analysis since the number of rare diseases is itself small.
2. It enhances the privacy protection of published datasets in individual organizations. When all the data records in an organization are published, an adversary knows that a victim's data record is definitely in the dataset. When some data records are withheld (in this model, they are moved to centralized collection and publication), the adversary will have a reduced confidence that the record of the victim is in the dataset, and such a reduced confidence is good for protecting privacy from composition attacks.

3. It improves the data utility. First, after removing the low frequency sensitive values, a low level data generalization will protect the sensitive information of individuals because sensitive values with moderate to high frequency are likely appear in many equivalence classes such that the risk of composition attacks is small. When rare diseases are moved to a central data repository, their corresponding statistics will be more accurate than those from aggregations over individual datasets because the duplications are minimized. In turn the analysis will be more reliable on the centralized collection of rare diseases.

In sum, the risk of composition attack is greatly mitigated by the mixed publication model and the data utility is significantly improved. The centralization of a small portion of rare disease does not involve much additional cost for data management, and in many cases, this has been done by various medical registers.

8.7 Conclusion

Existing single data publication and serial data publication methods do not support multiple independent data publication by different publishers that share overlapping records. This chapter has discussed the cause of composition attack and the challenges for mitigating the risk. This chapter further analyzed two typical methods for mitigating the risk of composition attack and their strengths and weaknesses. An experimental comparison has been reported to show the strengths and weaknesses of both methods. Finally, we introduced a mixed publication model, combining local independent data publications with a global centralized publication, which greatly reduces the risk of composition attacks and maintains the utility of datasets.

Acknowledgements The work has been partially supported by Australian Research Council (ARC) Discovery Grant DP110103142 and a CORE (junior track) grant from the National Research Fund, Luxembourg.

Appendix

In this appendix, we discuss three measures used in the experiments and the definitions of differential privacy.

A. Metrics

Definition 8.1 (Kullback-Leibler Divergence). Kullback-Leibler (KL) divergence [19] is a non-symmetric measure of the difference between two probability distributions P and Q . Specifically, KL-divergence is a measure of information loss

when Q is used to approximate P . It is denoted by $D_{KL}(P||Q)$. If P and Q represent discrete probability distributions, KL-divergence of Q from P is defined as:

$$D_{KL}(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}$$

In other words, it is the expectation of the logarithmic difference between the probabilities P and Q , where the expectation is taken the probabilities P .

Definition 8.2 (City Block Distance). City block distance measures the similarity between two objects. If a and b are two objects described by a m -dimensional vector, then the city block distance between a and b is calculated as follows.

$$\text{Distance}(a, b) = \sum_{j=1}^m |a_j - b_j|$$

The city block distance is greater than or equal to zero. The distance is zero for identical objects and high for objects that share little similarity.

Definition 8.3 (Relative Error). Relative error indicates how accurate a measurement is relative to the actual value of an object being measured. If R_{act} represents the actual value and R_{est} represents the estimated value, then the relative error is defined as follows:

$$\text{Error} = \frac{|R_{act} - R_{est}|}{R_{act}} = \frac{\Delta R}{R_{act}}$$

where ΔR represents the absolute error.

B. Differential Privacy

Differential privacy has received significant attention recently because it provides semantic and cryptographically strong guarantees [18]. It ensures that an adversary learns little more about an individual being in the dataset than not [9, 18, 26].

“Differential privacy will ensure that the ability of an adversary to inflict harm (or good, for that matter) of any sort, to any set of people, should be essentially the same, independent of whether any individual opts in to, or opts out of, the dataset. [9]”

The intuition behind this is that the output from a differentially private mechanism is insensitive to any particular record.

Definition 8.4 (Differential Privacy [26]). A randomized function K is differentially private if for all datasets D and D' where their symmetric difference contains at most one record (i.e., $|D \Delta D'| \leq 1$), and for all possible anonymized dataset \hat{D} ,

$$\Pr[K(D) = \hat{D}] \leq e^\epsilon \times \Pr[K(D') = \hat{D}]$$

where the probabilities are over the randomness of K .

The parameter $\epsilon > 0$ is public and set by the data publishers [9]. The lower the value of ϵ , the stronger the privacy guarantee, whereas a higher value of ϵ provides more data utility [27]. Therefore, it is crucial to choose an appropriate value for ϵ to balance data privacy and utility [30, 33, 40].

A standard mechanism to achieve differential privacy is to add random noise to the original output of a function. The added random noise is calibrated according to the *sensitivity* of the function. The sensitivity of the function is the maximum difference between the values that the function may take on a pair of datasets that differ in only one record [9].

Definition 8.5 (Sensitivity [12]). For any function $f : D \rightarrow \mathbf{R}^d$, the sensitivity of f is measured as:

$$\Delta f = \max_{D, D'} ||f(D) - f(D')||_1$$

for all D, D' differing in at most one record.

Dwork et al. [12] suggest the Laplace mechanism to achieve differential privacy. The Laplace mechanism takes a dataset D , a function f and the parameter b to generate noise according to the Laplace distribution with the probability density function $\Pr(x|b) = \frac{1}{2b} \exp(-|x|/b)$ with variance $2b^2$ and mean 0. Theorem 8.1 connects the sensitivity to the magnitude of the noise that generates the noisy output $f(\hat{D}) = f(D) + \text{Lap}(b)$ to satisfy ϵ -differential privacy. Note that $\text{Lap}(b)$ is a random variable sampled from the Laplace distribution.

Theorem 8.1 ([10]). For any function $f : D \rightarrow \mathbf{R}^d$, the randomized function K that adds independently generated noise with distribution $\text{Lap}(\Delta/\epsilon)$ to each of the outputs guarantees ϵ -differential privacy.

Therefore, the function f , returns the count value, first computes the original count $f(D)$ and then outputs the noisy answer $f(\hat{D}) = f(D) + \text{Lap}(1/\epsilon)$.

References

1. Baig, M.M., Li, J., Liu, J., Ding, X., Wang, H.: Data privacy against composition attack. In: Proceedings of the 17th International Conference on Database Systems for Advanced Applications, pp. 320–334, Busan (2012)
2. Baig, M.M., Li, J., Liu, J., Wang, H.: Cloning for privacy protection in multiple independent data publications. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 885–894, Glasgow (2011)

3. Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar, K.: Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In: Proceedings of the 26th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 273–282, Beijing (2007)
4. Bu, Y., Fu, A.W., Wong, R.C.W., Chen, L., Li, J.: Privacy preserving serial data publishing by role composition. Proc. VLDB Endowment **1**(1), 845–856 (2008)
5. Cebul, R.D., Rebitzer, J.B., Taylor, L.J., Votruba, M.: Organizational fragmentation and care quality in the U.S. health care system. J. Econ. Perspect. (2008). doi:[10.3386/w14212](https://doi.org/10.3386/w14212)
6. Collaboration, P.S.: Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. Lancet **360**(9349), 1903–1913 (2002)
7. Cormode, G., Procopiuc, C.M., Shen, E., Srivastava, D., Yu, T.: Empirical privacy and empirical utility of anonymized data. In: Proceedings of the 29th IEEE International Conference on Data Engineering Workshops, pp. 77–82, Brisbane (2013)
8. Dwork, C.: Differential privacy. In: Proceedings of the 5th International Colloquium on Automata, Languages and Programming, pp. 1–12, Venice (2006)
9. Dwork, C.: A firm foundation for private data analysis. Commun. ACM **54**(1), 86–95 (2011)
10. Dwork, C., Kenthapadi, K., Mcsherry, F., Mironov, I., Naor, M.: Our data , ourselves : privacy via distributed noise generation. In: Advances in Cryptology - EUROCRYPT, pp. 486–503, St. Petersburg (2006)
11. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Proceedings of the 3rd Conference on Theory of Cryptography, pp. 265–284, Berlin (2006)
12. Dwork, C., Smith, A.: Differential privacy for statistics : what we know and what we want to learn. J. Priv. Confidentiality **1**(2), 135–154 (2009)
13. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: a survey of recent developments. ACM Comput. Surv. **42**(4), 1–53 (2010)
14. Fung, B.C.M., Wang, K., Fu, A.W., Pei, J.: Anonymity for continuous data publishing. In: Proceeding of the 11th International Conference on Extending Database Technology, pp. 264–275, Nantes (2008)
15. Ganta, S.R., Prasad, S., Smith, A.: Composition attacks and auxiliary information in data. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 265–273, Las Vegas, Nevada (2008)
16. Jiang, W., Clifton, C.: A secure distributed framework for achieving k-anonymity. VLDB J. **15**(4), 316–333 (2006)
17. Jurczyk, P., Xiong, L.: Towards privacy-preserving integration of distributed heterogeneous data. In: Proceedings of the 2nd Ph.D. Workshop on Information and Knowledge Management, pp. 65–72, Napa Valley, California (2008)
18. Kasiviswanathan, S.P., Smith, A.: On the ‘semantics’ of differential privacy: a Bayesian formulation. Priv. Confidentiality **6**(1), 1–16 (2014)
19. Kullback, S., Leibler, R.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
20. LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S., Kruschwitz, N.: Big data, analytics, and the path from insights to value. MIT Sloan Manag. Rev. **52**, 21–31 (2011)
21. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k -anonymity. In: Proceedings of the 22nd IEEE International Conference on Data Engineering, pp. 25–25, Atlanta, Georgia (2006)
22. Li, N., Li, T., Venkatasubramanian, S.: t -closeness: privacy beyond k -anonymity and l -diversity. In: Proceedings of the 23rd IEEE International Conference on Data Engineering, pp. 106–115, Istanbul (2007)
23. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l -diversity: privacy beyond k -anonymity. ACM Trans. Knowl. Discov. Data **1**(1) (2007). doi: [10.1145/1217299.1217302](https://doi.org/10.1145/1217299.1217302)

24. Malin, B., Sweeney, L.: How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *J. Biomed. Inform.* **37**, 179–192 (2004)
25. Mohammed, N., Fung, B.C.M., Wang, K., Hung, P.C.K.: Privacy-preserving data mashup. In: Proceeding of the 12th International Conference on Extending Database Technology, pp. 228–239, Saint Petersburg (2009)
26. Mohammed, N., Chen, R., Fung, B.C., Yu, P.S.: Differentially private data release for data mining. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 493–501, San Diego, California (2011)
27. Muralidhar, K., Sarathy, R.: Does differential privacy protect Terry Gross privacy? In: Domingo-Ferrer, J., Magkos, E., (eds.) *Privacy in Statistical Databases*. Lecture Notes in Computer Science, vol. 6344, pp. 200–209. Springer, Berlin (2010)
28. Newton, K.M., Peissig, P.L., Kho, A.N., Bielinski, S.J., Berg, R.L., Choudhary, V., Basford, M., Chute, C.G., Kullo, I.J., Li, R., Pacheco, J.A., Rasmussen, L.V., Spangler, L., Denny, J.C.: Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J. Am. Med. Inform. Assoc.* **20**(e1), e147–e154 (2013)
29. Provost, F., Fawcett, T.: Data science and its relationship to big data and data-driven decision making. *Big Data* **1**(1), 51–59 (2013)
30. Sarathy, R., Muralidhar, K.: Evaluating Laplace noise addition to satisfy differential privacy for numeric data. *Trans. Data Priv.* **4**(1), 1–17 (2011)
31. Sattar, S.A., Li, J., Ding, X., Liu, J., Vincent, M.: A general framework for privacy preserving data publishing. *Knowl.-Based Syst.* **54**(0), 276–287 (2013)
32. Sattar, S.A., Li, J., Liu, J., Heatherly, R., Malin, B.: A probabilistic approach to mitigate composition attacks on privacy in non-coordinated environments. *Knowl.-Based Syst.* **67**(0), 361–372 (2014)
33. Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., Martínez, S.: Enhancing data utility in differential privacy via microaggregation-based k -anonymity. *VLDB J.* **23**(5), 771–794 (2014)
34. Sweeney, L.: k -anonymity: A model for protecting privacy. *Int. J. Uncertainty Fuzziness Knowledge Based Syst.* **10**(5), 557–570 (2002)
35. Tene, O., Polonetsky, J.: Privacy in the age of big data: a time for big decisions. *Stanford Law Rev.* **64**, 63–69 (2012)
36. Wang, K., Fung, B.C.M.: Anonymizing sequential releases. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 414–423, Philadelphia, PA (2006)
37. Wong, R.C., Li, J., Fu, A.W., Wang, K.: (α, k) -anonymity: an enhanced k -anonymity model for privacy preserving data publishing. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 754–759, Philadelphia, PA (2006)
38. Wong, R.C., Fu, A.W., Liu, J., Wang, K., Xu, Y.: Global privacy guarantee in serial data publishing. In: Proceedings of 26th IEEE International Conference on Data Engineering, pp. 956–959, Long Beach, California (2010)
39. Xiao, X., Tao, Y.: m -invariance: towards privacy preserving re-publication of dynamic data sets. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 689–700, Beijing (2007)
40. Xiao, X., Wang, G., Gehrke, Gehrke, J., Jefferson, T.: Differential privacy via wavelet transforms. *IEEE Trans. Knowl. Data Eng.* **23**(8), 1200–1214 (2011)
41. Xiong, L., Sunderam, V., Fan, L., Goryczka, S., Pourmajaf, L.: Predict: privacy and security enhancing dynamic information collection and monitoring. *Procedia Comput. Sci.* **18**(0), 1979–1988 (2013)

Chapter 9

Statistical Disclosure Limitation for Health Data: A Statistical Agency Perspective

Natalie Shlomo

Abstract Statistical agencies release health data collected in surveys, censuses and registers. In this chapter, statistical disclosure limitation (SDL) from the perspective of statistical agencies is presented. Traditional outputs in the form of survey microdata and tabular outputs are first presented with respect to quantifying disclosure risk, common SDL techniques for protecting the data, and measuring information loss. In recent years, however, there is greater demand for data including government ‘open data’ initiatives, which have led statistical agencies to examine additional forms of disclosure risks, related to the concept of differential privacy in the computer science literature. A discussion on whether SDL practices carried out at statistical agencies for traditional outputs are differentially private, is provided in the chapter. The chapter concludes with the presentation of some innovative data dissemination strategies that are currently being assessed by statistical agencies, where stricter privacy guarantees are necessary.

9.1 Introduction

Statistical agencies have an obligation to release statistical data for research purposes and to inform policies. On the other hand, they have a legal, moral and ethical commitment to protect the confidentiality of individuals responding to their request for data via surveys, censuses and registers. Health data that are collected by statistical agencies are particularly sensitive since they may contain sensitive information, such as diagnosis codes and abortion statistics.

Statistical agencies generally have two approaches for statistical disclosure limitation (SDL): *protect the data for release using disclosure limitation techniques* (“safe data”), or *restrict access to the data*, for example by limiting its use to approved researchers within a secure data environment (“safe access”). When releasing sensitive health data, a combination of both approaches is usually applied.

N. Shlomo (✉)

Social Statistics, School of Social Sciences, University of Manchester, Manchester, UK

e-mail: natalie.shlomo@manchester.ac.uk

Statistical data that are traditionally released by statistical agencies include microdata from social surveys, such as the General Health Survey, and tabular data. Tabular data contain either frequency counts or magnitude data which typically arise from business surveys, e.g., total revenue by industry code. For each of these traditional outputs, there has been much research on how to quantify disclosure risk, which SDL techniques to use to protect the data and their impact on information loss. However, with increasing demands for new forms of data at higher resolution, in particular linked hierarchical data and open data initiatives, there is even more pressure on statistical agencies to broaden access and to provide better solutions for the release of statistical data.

In traditional outputs, we define the notion of an *intruder*, as someone who wants to attack statistical data for malicious intent. In terms of health data, intruders might be, for example, individuals or organizations who wish to disclose sensitive information about doctors/clinics performing abortions. Two main disclosure risks are *identity disclosure*, where a statistical unit can be identified based on a set of cross-classified identifying variables, such as age, gender, occupation, place of residence; and *attribute disclosure*, where new information can be learnt about an individual or a group of individuals. Disclosure risk scenarios form the basis of possible means of disclosure, for example, the ability of an intruder to match a dataset to an external public file based on a common set of identifying variables; the ability of an intruder to identify uniques through visible and rare attributes; the ability of an intruder to difference nested frequency tables and obtain small counts; the ability of an intruder to form coalitions, and so on.

In microdata from social surveys, the main concern is the *risk of identification*, since it is a prerequisite for attribute disclosure where many new sensitive variables, such as income or health outcomes, may be revealed following an identification. Naturally, sampling from the population provides a-priori protection, since an intruder cannot be certain whether a sample unique is a population unique. In tabular data of whole population counts, attribute disclosure arises when there is a row / column of zeros and only one populated cell. This leads to individual or group attribute disclosure depending on the size of the populated cell, since an intruder can learn new attributes based on the remaining spanning variables of the table that was not known previously. Therefore, in frequency tables containing whole population counts, it is the zero cells that are the main cause of attribute disclosure. Frequency tables of weighted survey counts are generally not a cause of concern due to the ambiguity introduced by the sampling and the survey weights. In magnitude tables arising from business surveys, disclosure risk is generally defined by the ability of businesses to be able to form coalitions to disclose sensitive information about competing businesses. Since our focus is on health data arising from official statistics, the SDL of business data will not be considered further.

One disclosure risk that is often overlooked in traditional statistical outputs and is only now coming to prominence with ongoing research and development into web-based interactive data dissemination is *inferential disclosure*. Inferential disclosure risk is the ability to learn new attributes with high probability. For example, a proportion of some characteristic is very high within a subgroup, e.g.,

a high proportion of those who smoke have heart disease, or a regression model may have very high predictive power if the dependent and explanatory variables are highly correlated, e.g., regressing BMI on height and weight. In fact, an individual does not even have to be in the dataset in order to disclose information. Another example of inferential disclosure is *disclosure by differencing frequency tables* of whole population counts, when multiple tabular releases are disseminated from one data source. Inferential disclosure risk forms the basis of the definition of *differential privacy* as formulated in the computer science literature. Although the theory of differential privacy was developed in the context of the protection of queries in a remote query system, it is now coming to the attention of statistical agencies who are considering online interactive dissemination of statistical data and therefore are in need of stricter privacy guarantees.

An overview of statistical disclosure limitation for traditional outputs: microdata from social surveys which include the General Health Survey and whole population tabular data, are presented in Sects. 9.2 and 9.3 respectively. In Sect. 9.4, we discuss differential privacy and inferential disclosure as developed in the computer science literature and consider these definitions in the context of survey sampling and perturbation techniques, as practiced at statistical agencies. In Sect. 9.5 we present a future outlook for releasing statistical data. We conclude in Sect. 9.6 with some considerations for the future of data dissemination at statistical agencies.

9.2 Statistical Disclosure Limitation for Microdata from Social Surveys

Disclosure risk typically arises from attribute disclosure where small counts on cross-classified indirectly identifying key variables, such as: age, sex, place of residence, marital status, occupation, can be used to identify an individual and confidential information may be learnt. Identity disclosure therefore is a prerequisite for attribute disclosure and disclosure risk assessment is based on estimating a probability of identification given a set of identifying key variables. Generally, identifying key variables are categorical. Sensitive variables are often continuous, but can also be categorical. In order to protect the data, one can either apply an SDL technique on the identifying key variables or the sensitive variables. In the first case, identification of a unit is rendered more difficult, and the probability that a unit is identified is hence reduced. In the second case, even if an intruder succeeds in identifying a unit by using the values of the identifying key variables, the sensitive variables would hardly disclose any useful information on the particular record. One can also apply SDL techniques on both the identifying and sensitive variables simultaneously. This offers more protection, but also leads to more information loss.

In Sect. 9.2.1, we review the probabilistic modelling approach to disclosure risk assessment (see [4, 5, 16, 42, 43] and references therein). The probabilistic models provide consistent global and record-level disclosure risk measures, which

is an important feature for informing decisions about data release and where to target additional SDL techniques if necessary. The probabilistic modelling takes into account the protection afforded by the sampling and is based on the notion of population uniqueness. Shlomo and Skinner [37] have further developed this approach to take into account the realistic case where identifying variables may be misclassified or purposely perturbed. In Sect. 9.2.2, we present two examples of perturbative SDL methods that have been adapted to preserve sufficient statistics, as well as logical consistencies of the data. Subsequently, in Sect. 9.2.3 we discuss some common information loss measures.

9.2.1 Disclosure Risk Assessment

Identifying key variables for disclosure risk assessment are determined by a disclosure risk scenario, i.e., assumptions about available external files and IT tools that can be used by intruders to identify individuals in released microdata. For example, key variables may be chosen, which would enable matching the released microdata to a publicly available file containing names and addresses. Examples of publicly available data might include data that is freely available over the internet, such as car registrations, phone book or electoral roles, or can be purchased, such as supermarket loyalty cards and life-style datasets.

Under a probabilistic approach, disclosure risk is assessed on the contingency table of counts spanned by these identifying key variables. The other variables in the file are sensitive variables. The assumption is that the microdata contain individuals investigated in a survey and the population is unknown (or only partially known through some marginal distributions). The disclosure risk is a function of both the population and the sample, and in particular the cell counts of a contingency table defined by combinations of identifying discrete key variables, for example place of residence, sex, age and occupation.

Individual per-record risk measures, in the form of a probability of re-identification, are estimated. These per-record risk measures are then aggregated to obtain global risk measures for the entire file. Denoting by F_k the population size in cell k of a table spanned by key variables having K cells; f_k the sample size, $\sum_{k=1}^K F_k = N$ and $\sum_{k=1}^K f_k = n$, then the set of sample uniques is defined as $SU = \{k : f_k = 1\}$, since these are potential high-risk records, i.e., population uniques. Two global disclosure risk measures (where I is the indicator function) are the following:

1. Number of sample uniques that are population uniques:

$$\tau_1 = \sum_k I(f_k = 1, F_k = 1)$$

2. Expected number of correct matches for sample uniques (i.e., a matching probability): $\tau_2 = \sum_k I(f_k = 1) 1/F_k$.

The individual risk measure for τ_2 is $1/F_k$. This is the probability that a match between a record in the microdata and a record in the population having the same

values of key variables will be correct. If, for example, there are two records in the population with the same values of key variables, the probability is 0.5 that the match will be correct. Adding up these probabilities over the sample uniques gives the expected number (on average) of correctly matching a record in the microdata to the population when we allow guessing. The population frequencies are unknown and need to be estimated from the probabilistic model. The disclosure risk measures are estimated by:

$$\hat{\tau}_1 = \sum_k I(f_k = 1) \hat{P}(F_k = 1 | f_k = 1), \quad \hat{\tau}_2 = \sum_k I(f_k = 1) \hat{E}(1/F_k | f_k = 1) \quad (9.1)$$

Skinner and Holmes [42] and Elamir and Skinner [16] propose a Poisson Model to estimate disclosure risk measures. In this model, they make the natural assumption in contingency table literature: $F_k \sim \text{Poisson}(\lambda_k)$ for each cell k . A sample is drawn by Poisson or Bernoulli sampling with a sampling fraction π_k in cell k : $f_k | F_k \sim \text{Bin}(F_k, \pi_k)$. It follows that:

$$f_k \sim \text{Poisson}(\pi_k \lambda_k) \text{ and } F_k | f_k \sim \text{Poisson}(\lambda_k(1 - \pi_k)) \quad (9.2)$$

where $F_k | f_k$ are conditionally independent.

The parameters $\{\lambda_k\}$ are estimated using log-linear modeling. The sample frequencies f_k are independent Poisson distributed with a mean of $\mu_k = \pi_k \lambda_k$. A log-linear model for the μ_k is expressed as: $\log(\mu_k) = x'_k \beta$, where x_k is a design vector which denotes the main effects and interactions of the model for the key variables. The maximum likelihood estimator (MLE) $\hat{\beta}$ may be obtained by solving the score equations:

$$\sum_k [f_k - \pi_k \exp(x'_k \beta)] x_k = 0 \quad (9.3)$$

The fitted values are calculated by $\hat{u}_k = \exp(x'_k \hat{\beta})$ and $\hat{\lambda}_k = \hat{u}_k / \pi_k$. Individual disclosure risk measures for cell k are:

$$\begin{aligned} P(F_k = 1 | f_k = 1) &= \exp(\lambda_k(1 - \pi_k)), \\ E(1/F_k | f_k = 1) &= [1 - \exp(\lambda_k(1 - \pi_k))] / [\lambda_k(1 - \pi_k)] \end{aligned} \quad (9.4)$$

Plugging $\hat{\lambda}_k$ for λ_k in (9.4) leads to the estimates $\hat{P}(F_k = 1 | f_k = 1)$ and $\hat{E}[1/F_k | f_k = 1]$ and then to $\hat{\tau}_1$ and $\hat{\tau}_2$ of (9.1). Rinott and Shlomo [32] consider confidence intervals for these global risk measures.

Skinner and Shlomo [43] develop a method for selecting the log-linear model based on estimating and (approximately) minimizing the bias of the risk estimates $\hat{\tau}_1$ and $\hat{\tau}_2$. Defining $h(\lambda_k) = P(F_k = 1 | f_k = 1)$ for τ_1 and $h(\lambda_k) = E(1/F_k | f_k = 1)$ for τ_2 , they consider the expression: $B = \sum_k E[I(f_k = 1)][h(\hat{\lambda}_k) - h(\lambda_k)]$. A Taylor expansion of h leads to the approximation:

$$B \approx \sum_k \pi_k \lambda_k \exp(-\lambda_k) [h'(\lambda_k)(\hat{\lambda}_k - \lambda_k) + h''(\lambda_k)(\hat{\lambda}_k - \lambda_k)^2/2]$$

and the relations $E f_k = \pi_k \lambda_k$ and $E[(f_k - \pi_k \hat{\lambda}_k)^2 - f_k] = \pi_k^2 E(\lambda_k - \hat{\lambda}_k)^2$ under the hypothesis of a Poisson fit, lead to a further approximation of B of the form:

$$B \approx \sum_k \hat{\lambda}_k \exp(-\pi_k \hat{\lambda}_k) [-h'(\hat{\lambda}_k)(f_k - \pi_k \hat{\lambda}_k) + h''(\hat{\lambda}_k)[(f_k - \pi_k \hat{\lambda}_k)^2 - f_k]/(2\pi_k)] \quad (9.5)$$

For example, for τ_1 they obtain:

$$\hat{B}_1 \approx \sum_k \hat{\lambda}_k \exp(-\hat{\lambda}_k) (1 - \pi_k) \{(f_k - \pi_k \hat{\lambda}_k) + (1 - \pi_k)[(f_k - \pi_k \hat{\lambda}_k)^2 - f_k]/(2\pi_k)\} \quad (9.6)$$

The method selects the model using a forward search algorithm which minimizes the standardized bias estimate $\hat{B}_i/\sqrt{\hat{v}_i}$ for \hat{t}_i , with $i = 1, 2$, where \hat{v}_i are variance estimates of \hat{B}_i .

Empirical studies have shown that the probabilistic modelling approach can provide unbiased estimates of the overall global level of disclosure risk in the microdata, but are not accurate for the individual record level of disclosure risk. Thus care should be taken when using record level measures of disclosure risk for targeting SDL techniques to high-risk records.

Skinner and Shlomo [43] also address the estimation of disclosure risk measures under complex survey designs with stratification, clustering and survey weights. While the method described assumes that all individuals within cell k are selected independently using Bernoulli sampling, i.e., $P(f_k = 1|F_k) = F_k \pi_k (1 - \pi_k)^{F_k-1}$, this may not be the case when sampling clusters (households). In practice, key variables typically include variables such as age, sex and occupation, that tend to cut across clusters. Therefore, the above assumption holds in practice in most household surveys and does not cause bias in the estimation of the risk measures. Inclusion probabilities may vary across strata, the most common stratification is on geography. Strata indicators should always be included in the key variables to take into account differential inclusion probabilities in the model. Under complex sampling, the $\{\lambda_k\}$ can be estimated consistently using pseudo-maximum likelihood estimation, where the estimating equation in (9.3) is modified as:

$$\sum_k [\hat{F}_k - \exp(x'_k \beta)] x_k = 0 \quad (9.7)$$

and \hat{F}_k is obtained by summing the survey weights in cell k : $\hat{F}_k = \sum_{i \in k} w_i$.

The resulting estimates $\{\hat{\lambda}_k\}$ are then plugged into expressions in (9.4) and π_k is replaced by the estimate $\hat{\pi}_k = f_k/F_k$. Note that the risk measures in (9.4) only depend on sample uniques and that the value of $\hat{\pi}_k$ in this case is simply the reciprocal of the survey weight. The test criteria \hat{B} is also adapted via the pseudo-maximum likelihood method.

The presented probabilistic model, as well as other probabilistic methods (see Bethlehem et al. [5], Benedetti et al. [4], Rinott and Shlomo [30, 31]), assume that there is no measurement error in the way the data is recorded. Besides typical errors in data capture, key variables can also be purposely misclassified as a means of masking the data. Shlomo and Skinner [37] adapt the estimation of risk measures to take into account measurement errors. Denoting the cross-classified key variables in the population and the microdata as X , and assuming that X in the microdata have undergone some misclassification or perturbation error, denoted by the value \tilde{X} , and determined independently by a misclassification matrix M ,

$$M_{kj} = P(\tilde{X} = k | X = j) \quad (9.8)$$

the record-level disclosure risk measure of a match with a sample unique under measurement error is:

$$\frac{M_{kk}(1 - \pi M_{kk})}{\sum_j F_j M_{kj}/(1 - \pi M_{kj})} \leq \frac{1}{F_k} \quad (9.9)$$

Under assumptions of small sampling fractions and small misclassification errors, the measure can be approximated by: $M_{kk}/\sum_j F_j M_{kj}$ or M_{kk}/\tilde{F}_k , where \tilde{F}_k is the population count with $\tilde{X} = k$. Aggregating the per-record disclosure risk measures, the global risk measure is:

$$\tau_2 = \sum_k I(f_k = 1) M_{kk}/\tilde{F}_k \quad (9.10)$$

Note that to calculate the measure only the diagonal of the misclassification matrix needs to be known, i.e., the probabilities of not being perturbed. Population counts are generally not known so the estimate in (9.10) can be obtained by probabilistic modeling with log-linear models as described above on the misclassified sample:

$$\hat{\tau}_2 = \sum_k I(\tilde{f}_k = 1) M_{kk} \hat{E}(1/\tilde{F}_k | \tilde{f}_k) \quad (9.11)$$

9.2.2 Statistical Disclosure Limitation Methods

Based on the disclosure risk assessment, statistical agencies must choose appropriate SDL methods either by perturbing, modifying, or summarizing the data. The choice depends on the mode of access, the requirements of the users and the impact on data utility. Choosing an optimal SDL method is an iterative process where a balance must be found between managing disclosure risk and preserving the utility in the microdata.

SDL methods for microdata include perturbative methods that alter the data and non-perturbative methods which limit the amount of information released. Examples of non-perturbative SDL methods that are often applied on social survey microdata are *global recoding* and *suppression* of values or whole key variables. *Sub-sampling records* is also a non-perturbative method and is often used for producing census microdata. Perturbative methods for masking continuous sensitive variables include: *adding random noise* (see Kim [23], Fuller [19], and Brand [6]); *micro-aggregation*, where records are grouped and their values are replaced by their average (see Defays and Nanopoulos [11], Anwar [3], Domingo-Ferrer and Mateo-Sanz [14]); *rounding* to a pre-selected rounding base; and *rank swapping*, where values between pairs of record within a small group are swapped (see Dalenius and Reiss [9], Fienberg and McIntyre [17]). Perturbative methods for categorical key variables include *record swapping* (typically swapping geography variables) and a more general *post-randomization* probability mechanism (PRAM), where categories of variables are changed or not changed according to a prescribed probability matrix and a stochastic selection process (see Gourweleeuw et al. [21]). For more information on perturbative and non-perturbative methods see also: Willenborg and De Waal [44], Domingo-Ferrer, et al. [13], and references therein.

Each SDL method impacts differently the level of protection obtained in the microdata and information loss. Shlomo and De Waal [38] discuss optimizing SDL methods to preserve sufficient statistics, as well as logical consistencies in the microdata. Two of these perturbative techniques are described below.

9.2.2.1 PRAM for Categorical Key Variables

For protecting categorical identifying variables, Gourweleeuw et al. [21] propose the post-randomization method (PRAM). As a perturbative method, PRAM alters the data, and therefore we can expect consistent records to start failing edit rules. Edit rules describe logical relationships that have to hold true, such as “a ten-year old person cannot be married” or “the profit and the costs of an enterprise should sum up to its turnover”.

Willenborg and De Waal [44] describe the process of applying PRAM as follows: Let \mathbf{P} be a $L \times L$ transition matrix containing conditional probabilities $p_{ij} = p(\text{perturbed category is } j | \text{original category is } i)$ for a categorical variable with L categories, \mathbf{t} the vector of frequencies and \mathbf{v} the vector of relative frequencies: $\mathbf{v} = \mathbf{t}/n$, where n is the number of records in the micro-data set. In each record of the data set, the category of the variable is changed or not changed according to the prescribed transition probabilities in the matrix \mathbf{P} and the result of a draw of a random multinomial variate u with parameters p_{ij} ($j = 1, \dots, L$). If the j -th category is selected, category i is moved to category j . When $i = j$, no change occurs. Let \mathbf{t}^* be the vector of the perturbed frequencies. \mathbf{t}^* is a random variable and $E(\mathbf{t}^* | \mathbf{t}) = \mathbf{t}\mathbf{P}$. Assuming that the transition probability matrix \mathbf{P} has an inverse \mathbf{P}^{-1} , this can be used to obtain an unbiased moment estimator of the original data: $\hat{\mathbf{t}} = \mathbf{t}^*\mathbf{P}^{-1}$. In order to ensure that the transition probability matrix has an inverse

and to control the amount of perturbation, the matrix \mathbf{P} is chosen to be dominant on the main diagonal, i.e., each entry on the main diagonal is over 0.5.

The condition of invariance can be placed on the transition matrix \mathbf{P} , i.e., $\mathbf{tP} = \mathbf{t}$. This releases the users of the perturbed file of the extra effort to obtain unbiased moment estimates of the original data, since \mathbf{t}^* itself will be an unbiased estimate of \mathbf{t} . To obtain an invariant transition matrix, a matrix \mathbf{Q} is calculated by transposing matrix \mathbf{P} , multiplying each column j by v_j and then normalizing its rows so that the sum of each row equals one. The invariant matrix is obtained by $\mathbf{R} = \mathbf{PQ}$. The invariant matrix \mathbf{R} may distort the desired probabilities on the diagonal, so Shlomo and De Waal [38] define a parameter α and calculate $\mathbf{R}^* = \alpha\mathbf{R} + (1 - \alpha)\mathbf{I}$, where \mathbf{I} is the identity matrix. \mathbf{R}^* will also be invariant and the amount of perturbation is controlled by the value of α . The property of invariance means that the expected values of the marginal distribution of the variable being perturbed are preserved. In order to obtain the exact marginal distribution and reduce the additional variance caused by the perturbation, a “without” replacement selection strategy for choosing values to perturb can be implemented based on the expectations calculated from the transition probabilities.

As in most perturbative SDL methods, joint distributions between perturbed and unperturbed variables are distorted, in particular for variables that are highly correlated with each other. The perturbation can be controlled as follows:

1. Before applying PRAM, the variable to be perturbed is divided into subgroups, $g = 1, \dots, G$. The transition (and invariant) probability matrix is developed for each subgroup g , R_g . The transition matrices for each subgroup are placed on the main diagonal of the overall transition matrix where the off-diagonal probabilities are all zero, i.e., the variable is only perturbed within the subgroup and the difference in the variable between the original value and the perturbed value will not exceed a specified level. An example of this is perturbing *age* within broad age bands.
2. The variable to be perturbed may be highly correlated with other variables. Those variables should be compounded into one single variable. PRAM should be carried out on the compounded variable. Alternatively, the variable to be perturbed is carried out within subgroups defined by the second highly correlated variable. An example of this is when *age* is perturbed within groupings defined by *marital status*.

The control variables in the perturbation process will minimize the amount of edit failures, but they will not eliminate all edit failures, especially edit failures that are out of scope of the variables that are being perturbed. Remaining edit failures need to be manually or automatically corrected through edit and imputation processes depending on the amount and types of edit failures.

9.2.2.2 Additive Noise for Continuous Variables

In its basic form, random noise is generated independently and identically distributed with a positive variance and a mean of zero. The random noise is then added to the original variable (see Brand [6] and references therein for a summary and discussion of additive random noise). Adding random noise will not change the mean of the variable for large datasets, but will introduce more variance. This will impact on the ability to make statistical inferences. Researchers may have a suitable methodology to correct for this type of measurement error, but it is good practice to minimize these errors through better implementations of the method.

Additive noise should be generated within small homogenous sub-groups (for example, percentiles of the continuous variable) in order to use different initiating perturbation variance for each sub-group. Generating noise in sub-groups also causes less edit failures with respect to relationships in the data. Following Kim [23] and Fuller [19], correlated random noise can be added to the continuous variable thereby ensuring that not only means are preserved but also the exact variance. A simple method for generating correlated random noise for a continuous variable z , as described in Shlomo and De Waal [38], operates as follows:

Procedure 1 (Univariate). Define a parameter δ which takes a value greater than zero and less than or equal to one. When $\delta = 1$ we obtain the case of fully modelled synthetic data. The parameter δ controls the amount of random noise added to the variable z . After selecting a δ , calculate $d_1 = \sqrt{(1 - \delta^2)}$ and $d_2 = \sqrt{\delta^2}$. Now, generate random noise ε independently for each record with a mean of $\mu' = \frac{1-d_1}{d_2} \mu$ and the original variance of the variable σ^2 . Typically, a Normal Distribution is used to generate the random noise. Calculate the perturbed variable z'_i for each record i in the sample microdata ($i = 1, \dots, n$) as a linear combination: $z'_i = d_1 \times z_i + d_2 \times \varepsilon_i$. Note that $E(z') = d_1 E(z) + d_2 [\frac{1-d_1}{d_2} E(z)] = E(z)$ and $Var(z') = (1 - \delta^2) Var(z) + \delta^2 Var(z) = Var(z)$, since the random noise is generated independently to the original variable z .

An additional problem when adding random noise is that there may be several variables to perturb at once, and these variables may be connected through an edit constraint of additivity. One procedure to preserve additivity would be to perturb two of the variables and obtain the third from aggregating the perturbed variables. However, this method will not preserve the total, mean and variance of the aggregated variable and in general, it is not good practice to compound effects of perturbation by aggregating perturbed variables since this causes unnecessary information loss.

Shlomo and De Waal [38] propose implementing Procedure 1 in a multivariate setting, where correlated Gaussian noise is added to the variables simultaneously. The method not only preserves the means of each of the three variables and their co-variance matrix, but also preserves the edit constraint of additivity.

Procedure 2 (multivariate). Consider three variables x, y and z , where $x + y = z$. This procedure generates random noise that *a-priori* preserves additivity and

therefore combining the random noise to the original variables will also ensure additivity. In addition, means and the covariance structure are preserved. The technique works as follows. Generate multivariate random noise: $(\varepsilon_x, \varepsilon_y, \varepsilon_z)^T \sim N(\mu', \Sigma)$, where the superscript T denotes the transpose. In order to preserve sub-totals and limit the amount of noise, the random noise should be generated with percentiles (note that we drop the index for percentiles). The vector μ' contains the corrected means of each of the three variables x, y and z based on the noise parameter δ : $\mu'^T = (\mu'_x, \mu'_y, \mu'_z) = (\frac{1-d_1}{d_2} \mu_x, \frac{1-d_1}{d_2} \mu_y, \frac{1-d_1}{d_2} \mu_z)$. The matrix Σ is the original covariance matrix. For each separate variable, calculate the linear combination of the original variable and the random noise as previously described. For example, for record $i : z'_i = d_1 z_i + d_2 \varepsilon_{zi}$. The mean vector and the covariance matrix remain the same before and after the perturbation, and the additivity is exactly preserved.

9.2.3 Information Loss Measures

The information loss in microdata that has undergone SDL techniques is assessed on whether the same statistical analysis and inference can be drawn on the perturbed data compared to the original data. Microdata is multi-purposed and used by many different types of users, with diverse reasons for analysing the data. To assess the utility in microdata, proxy measures have been developed and include measuring distortions to distributions and the impact on bias, variance and other statistics (Chi-squared statistic, R2 goodness of fit, rankings, etc.). Domingo-Ferrer, et al. [13], Gomatam and Karr [20], Shlomo and Young [40, 41], and Shlomo [35], describe the use of such measures for assessing information loss in perturbed statistical data with empirical examples and applications. A brief summary of some useful proxy measures is presented below.

9.2.3.1 Distance Metrics

Distance metrics are used to measure distortions to distributions in the microdata as a result of applying SDL techniques. The AAD is a distance metric based on the average absolute difference between observed and perturbed counts in a frequency distribution. Let D represent a frequency distribution produced from the microdata and let $D(c)$ be the frequency in cell c . The Average Absolute Distance (AAD) per cell is defined as:

$$AAD(D_{orig}, D_{pert}) = \sum_c |D_{pert}(c) - D_{orig}(c)| / n_c \quad (9.12)$$

where n_c is the number of cells in the distribution.

9.2.3.2 Impact on Measures of Association

Tests for independence are often carried out on joint frequency distributions between categorical variables that span a table calculated from the microdata. The test for independence for a two-way table is based on a Pearson Chi-Squared Statistic $\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$, where o_{ij} is the observed count and $e_{ij} = (n_i n_j)/n$ is the expected count for row i and column j . If the row and column are independent then χ^2 has an asymptotic chi-square distribution with $(R - 1)(C - 1)$ and for large values the test rejects the null hypothesis in favor of the alternative hypothesis of association. Typically, the Cramer's V is used, which is a measure of association between two categorical variables: $CV = \sqrt{\frac{\chi^2/n}{\min\{(R-1), (C-1)\}}}$. The information loss measure is the percent relative difference between the original and perturbed table:

$$RCV(D_{pert}, D_{orig}) = 100 \times \frac{CV(D_{pert}) - CV(D_{orig})}{CV(D_{orig})} \quad (9.13)$$

For multiple dimensions, log-linear modeling is often used to examine associations. A similar measure to (9.13) can be calculated by taking the relative difference in the deviance obtained from the model based on the original and perturbed microdata.

9.2.3.3 Impact on Regression Analysis

For continuous variables, it is useful to assess the impact on the correlation and in particular the R^2 of a regression (or ANOVA) analysis. For example, in an ANOVA, the test involves whether a continuous dependent variable has the same means across groups defined by a categorical explanatory variable. The goodness of fit criterion R^2 is based on a decomposition of the variance of the mean of the dependent variable. By perturbing the statistical data, the groupings may lose their homogeneity, the “between” variance becomes smaller, and the “within” variance becomes larger. In other words, the proportions within each of the groupings shrink towards the overall mean. On the other hand, the “between” variance may become artificially larger showing more association than in the original distribution.

The utility is based on assessing differences in the means of a response variable across categories of an explanatory variable having K categories. Let \bar{y}_k be the mean in category k and define the “between” variance of this mean by $BV(\bar{y}_{orig}) = \frac{1}{K-1} \sum_k (\bar{y}_k - \bar{y})^2$, where \bar{y} is the overall mean. Information loss is measured by:

$$BVR(\bar{y}_{pert}, \bar{y}_{orig}) = 100 \times \frac{BV(\bar{y}_{pert}) - BV(\bar{y}_{orig})}{BV(\bar{y}_{orig})} \quad (9.14)$$

In addition, other analysis of information loss involves comparing estimates of coefficients when applying a regression model on both the original and perturbed microdata and comparing the coverage of confidence intervals.

9.3 Statistical Disclosure Limitation for Frequency Tables

The focus of this section is on protecting frequency tables of whole population counts. This is more challenging than protecting tables from a sample. The sampling a-priori introduces ambiguity into the frequency counts and, as a result, it is more difficult to identify statistical units without response knowledge or infer what the true count may be in the population. Moreover, tabular data from samples are typically weighted counts where sampling weights vary between units because of differential selection probabilities and non-response adjustments. Therefore, the number of contributors to a cell is not always known.

Tabular outputs are pre-determined by statistical agencies after careful consideration of population thresholds, average cell sizes, collapsing and fixing categories of variables spanning the tables. In spite of all these efforts, SDL techniques are still necessary. These techniques include pre-tabular methods, post-tabular methods and combinations of both.

Pre-tabular methods are implemented on the microdata prior to the tabulation of the tables. The most commonly used method is record swapping between a pair of households/individuals matching on some control variables (see Dalenius and Reis [9] and Fienberg and McIntyre [17]). This method has been used for protecting census tables at the United States Bureau of the Census and the Office for National Statistics (ONS) in the United Kingdom. Record swapping can be seen as a special case of the more general method of PRAM, as defined in Sect. 9.2.2.1. In practice, statistical agencies prefer record swapping since the method is easy to implement and marginal distributions are preserved exactly on higher aggregations of the data.

Post-tabular methods are implemented on the entries of the tables after they are computed and typically take the form of random rounding, either on the small cells of the tables or on all entries of the tables. Within the framework of developing the SDL software package, Tau Argus, a fully controlled rounding option has been added (Hundepool [22]). The procedure uses linear programming techniques to round entries up or down and in addition ensures that all rounded entries add up to the rounded totals. Other post-tabular methods include cell suppression or some form of random perturbation on the cells of the whole population tables.

In Sect. 9.3.1 we provide an overview of disclosure risks in tabular data and in Sect. 9.3.2 present disclosure risk and information loss measures based on Information Theory as developed in Antal et al. [2] and Shlomo et al. [34]. Section 9.3.3 presents some common SDL techniques.

9.3.1 Disclosure Risk in Whole Population Tabular Outputs

The main disclosure risk in a register/census context comes from small counts, i.e., ones and twos, since these can lead to re-identification. The amount and placement of the zeros in the table determines whether new information can be learnt about

an individual or a group of individuals. Therefore, SDL techniques for whole population tabular data should not only protect small cells in the tables but also introduce ambiguity and uncertainty into the zero values.

The disclosure risks as defined in [35, 44] are:

Individual attribute disclosure: An individual can be identified on the basis of some of the variables spanning the table and a new attribute revealed about the individual, i.e. for tabular data, this means that there is a one in a margin of the table. Identification is a necessary pre-condition for attribute disclosure and therefore should be avoided. In a census/register context where many tables are released, an identification made in a lower dimensional table will lead to attribute disclosure in a higher dimensional table.

Group attribute disclosure: If there is a row or column that contains mostly zeros and a small number of non-zero cells, then one can learn a new attribute about a group of individuals as well as learn about the group of individuals who do not have this attribute. This type of disclosure risk does not require individual identification.

Disclosure by differencing: Two tables that are nested may be subtracted one from the other resulting in a new table containing small cells and the above disclosure risk scenarios would apply. For example, a table containing the elderly population in private households may be subtracted from a table containing the total elderly population, resulting in a table of the elderly in communal establishments. This table is typically very sparse compared to the two original tables.

Disclosure by linking tables: Since many tables are disseminated from one data source, they can be linked through common cells and common margins thereby increasing the chances for revealing SDL methods and original cell counts.

To protect against attribute disclosure, SDL techniques should limit the risk of identification and also introduce ambiguity into the zero counts. To avoid disclosure by differencing, often only one set of variables and geographies are disseminated with no possibilities for overlapping categories. To avoid disclosure by linking tables, margins and cells of tables should be made consistent. In addition, statistical agencies often employ transparent and visible SDL techniques to avoid any perception that there may be disclosure risks in the data and resources are directed to ensure that the public is informed about the measures taken to protect confidentiality.

9.3.2 Disclosure Risk and Information Loss Measures Based on Information Theory

As mentioned, attribute disclosure is caused by rows/columns that have many zero cells and only one or two populated cells. A row/column with a uniform distribution of cell counts would have little attribute disclosure risk whilst a degenerate

distribution of cell counts would have high attribute disclosure risk. Moreover, a row/column with large counts would have less risk of re-identification compared to a row/column with small counts. Antal et al. [2] propose a new disclosure risk measure based on Information Theory.

In their paper, the disclosure risk measure for a whole-population frequency table should have the following properties: (a) small cell values have higher disclosure risk than large values; (b) uniformly distributed frequencies imply low disclosure risk; (c) the more zero cells in the census table, the higher the disclosure risk; (d) the risk measure should be bounded by 0 and 1. Using Information Theory, an analytical expression of disclosure risk that meets these properties can be computed. The entropy of the frequency vector in a table of size K , with population counts $F = (F_1, F_2, \dots, F_K)$, where $\sum_{i=1}^K F_i = N$ is:

$$H(P) = H\left(\frac{F}{N}\right) = -\sum_{i=1}^K \frac{F_i}{N} \log\left(\frac{F_i}{N}\right) = \frac{N \log N - \sum_{i=1}^K F_i \log F_i}{N}$$

and to produce a disclosure risk measure between zero and one, the disclosure risk measure is defined as:

$$1 - \frac{H\left(\frac{F}{N}\right)}{\log K} \quad (9.15)$$

The disclosure risk measure in (9.15) ensures property (b) since the term will tend to zero as the frequency distribution is more uniform, and ensures property (d) since the measure is bounded between zero and one. However, the disclosure risk measure does not take into account the magnitude of the cells counts or the number of zero cells in the table (or row/column of the table) and does not preserve properties (a) and (c). Therefore, an extended disclosure risk measure is proposed in (9.16) and is defined as a weighted average of three different terms, each term being a measure between zero and one.

$$R(F, w_1, w_2) = w_1\left[\frac{|A|}{K}\right] + w_2\left[1 - \frac{N \log N - \sum_{i=1}^K F_i \log F_i}{N \log K}\right] - (1 - w_1 - w_2)\left[\frac{1}{\sqrt{N}} \log \frac{1}{e\sqrt{N}}\right] \quad (9.16)$$

where A is the set of zeroes in the table and $|A|$ the number of zeros in the set, K , N and F as defined above, and w_1, w_2 are arbitrary weights: $0 \leq w_1 + w_2 \leq 1$.

The first measure in (9.16) is the proportion of zeros which is relevant for attribute disclosure and property (c). The third measure in (9.16) allows us to differentiate between tables with different magnitudes and accounts for property (a). As the population size N gets larger in the table, the third measure tends to zero. The weights w_1 and w_2 should be chosen depending on the data protector's choice of how important each of the terms are in contributing to disclosure risk. Alternatively,

one can avoid weights altogether by taking the L_2 norm of the three terms of the risk measure in (9.16) as follows:

$$\frac{(\sum_{i=1}^3 |x_i|^2)^{1/2}}{\sqrt{3}},$$

where x_i represents term i ($i = 1, 2, 3$) in (9.16). This provides more weight to the larger term of the risk measure.

Continuing with Information Theory based measures, Antal et al. [2] propose to measure information loss using Hellinger's Distance to assess the distance between two distributions. This is the L_2 norm:

$$HD(P, Q) = \frac{1}{\sqrt{2}} \|\sqrt{P} - \sqrt{Q}\|_2,$$

where $P = (p_1, p_2, \dots, p_k)$ is the original probability distribution of cell counts and Q is the perturbed probability distribution of cell counts.

The Hellinger's distance is bounded by zero and one, and preserves the properties of a distance metric (non-negativity, coincidence axiom, symmetry and triangle inequality). Measuring the distance infers that the smaller the distance the more information is left in the table.

In the case of frequency distributions from whole-population tables, where $F = (F_1, F_2, \dots, F_K)$ is the vector of original counts and $G = (G_1, G_2, \dots, G_K)$ is the vector of perturbed counts, and $\sum_{i=1}^K F_i = N$ and $\sum_{i=1}^K G_i = M$, the Hellinger Distance is defined as:

$$HD(F, G) = \frac{1}{\sqrt{2}} \|\sqrt{F} - \sqrt{G}\|_2 = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^K (\sqrt{F_i} - \sqrt{G_i})^2} \quad (9.17)$$

The Hellinger's Distance takes into account the magnitude of the cells since the difference between square roots of two "large" numbers is smaller than the difference between square roots of two "small" numbers, even if these pairs have the same absolute difference. The lower bound remains zero and the upper bound of this distance of counts changes:

$$\begin{aligned} HD(F, G) &= \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^K (\sqrt{F_i} - \sqrt{G_i})^2} = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^K (F_i + G_i - 2\sqrt{F_i G_i})} \\ &= \frac{1}{\sqrt{2}} \sqrt{N + M - 2 \sum_{i=1}^K \sqrt{F_i G_i}} \leq \sqrt{\frac{N + M}{2}} \end{aligned}$$

The information loss measure which is bounded by zero and one, where zero represents low utility and one represents high utility, is:

$$1 - \frac{HD(F, G)}{\sqrt{\frac{N+M}{2}}}$$

More details of these measures are in Shlomo et al. [34], which focuses on the development of flexible web-based table generation platforms for user-defined census tables and where the disclosure risk and information loss measures need to be calculated on-the-fly. In that paper, the disclosure risk measure is also adapted for assessing disclosure risk following the application of SDL techniques.

9.3.3 Statistical Disclosure Limitation Methods

In this section, we describe three SDL techniques which have been used to protect whole population frequency tables such as for a census or register: a pre-tabular SDL method of record swapping and post-tabular SDL methods of random rounding and stochastic perturbation. Record Swapping is used to protect census outputs in the United States and the United Kingdom, a post-tabular method of random rounding is used in New Zealand and Canada and a post-tabular probabilistic stochastic perturbation mechanism has been recently implemented in Australia. Examples of applications and case studies can be found in Shlomo [35], Shlomo and Young [40], and Shlomo et al. [34, 39].

9.3.3.1 Record Swapping

Record swapping is based on the exchange of values of variable(s) between similar pairs of population units (often households). In order to minimize bias, pairs of population units are determined within strata defined by control variables. For example, for swapping households, control variables may include: a large geographical area, household size and the age-sex distribution of individuals in the households. In addition, record swapping can be targeted to high-risk population units found in small cells of census tables. In a census/register context, geographical variables related to place of residence are often swapped. Swapping place of residence have the following properties: (1) it minimizes bias based on the assumption that place of residence is independent of other target variables conditional on the control variables; (2) it provides more protection in the tables since place of residence is a highly visible variable which can be used to identify individuals; (3) it preserves marginal distributions within a larger geographical area. For more information on record swapping, see Dalenius and Reiss [9], Fienberg and McIntyre [17], and Shlomo [35].

9.3.3.2 Semi-Controlled Random Rounding

Another post-tabular method of SDC for frequency tables is based on unbiased random rounding. Let $\text{Floor}(x)$ be the largest multiple bk of the base b , such that $bk < x$ for any value of x . In this case, $\text{res}(x) = x - \text{Floor}(x)$. For an unbiased rounding procedure, x is rounded up to $\text{Floor}(x) + b$ with probability $\text{res}(x)/b$ and rounded down to $\text{Floor}(x)$ with probability $(1 - \text{res}(x)/b)$. If x is already a multiple of b , it remains unchanged.

In general, each small cell is rounded independently in the table, i.e., a random uniform number u between zero and one is generated for each cell. If $u \leq \text{res}(x)/b$ then the entry is rounded up; otherwise, it is rounded down. This ensures an unbiased rounding scheme, i.e. the expectation of the rounding perturbation is zero. However, the realization of this stochastic process on a finite number of cells in a table will not ensure that the sum of the perturbations will exactly equal zero. To place some control in the random rounding procedure, we use a semi-controlled random rounding algorithm for selecting entries to round up or down as follows: First the expected number of entries of a given $\text{res}(x)$ that are to be rounded up is predetermined (for the entire table or for each row/column of the table). The expected number is rounded to the nearest integer. Based on this expected number, a random sample of entries is selected (without replacement) and rounded up. The other entries are rounded down. This process ensures that rounded internal cells aggregate to the controlled rounded total.

Due to the large number of perturbations in the table, margins are typically rounded separately from internal cells and tables are not additive. When using semi-controlled random rounding this alleviates some of the problems of non-additivity since one of the margins and the overall total will be controlled, i.e., the rounded internal cells aggregate to the rounded total.

Another problem with random rounding is the consistency of the rounding across same cells that are generated in different tables. It is important to ensure that the cell value is always rounded consistently, otherwise the true cell count can be learnt by generating many tables containing the same cell and observing the perturbation patterns. Fraser and Wooton [18] propose the use of *microdata keys* which can solve the consistency problem. First, a random number (which they call a *key*) is defined for each record in the microdata. When building a census frequency table, records in the microdata are combined to form a cell defined by the spanning variables of the table. When these records are combined to a cell, their keys are also aggregated. This aggregated key serves as the seed for the rounding and therefore same cells will always have the same seed and result in consistent rounding.

9.3.3.3 Stochastic Perturbation

A more general method than random rounding is stochastic perturbation, which involves perturbing the internal cells of a table using a probability transition matrix and is similar to the post-randomisation method (PRAM) that is used to perturb

categorical variables in microdata, as described in Sect. 9.2.2.1. In this case, it is the cell counts in a table that are perturbed. More details are in Fraser and Wooton [18] and Shlomo and Young [41].

In Sect. 9.2.2.1, we describe the invariant probability matrix \mathbf{R} . In this case \mathbf{t} is the vector of frequencies of the cell values where the last component would contain the number of cells above cap L and \mathbf{v} is the vector of relative frequencies: $\mathbf{v} = \mathbf{t}/K$, where K is the number of cells in the table. In each cell of the table, the cell value i is changed or not changed according to the prescribed transition probabilities in the invariant matrix \mathbf{R} . Placing the condition of invariance on the probability transition matrix means that the marginal distribution of the cell values are approximately preserved under the perturbation. As described in the random rounding procedure, in order to obtain the exact marginal distribution a similar strategy for selecting the cell values to change can be carried out. For each cell value i , the expected number of cells that need to be changed to a different value j is calculated according to the probabilities in the transition matrix. We then randomly select (without replacement) the expected number of cells i and carry out the change to j .

To preserve exact additivity in the table, an iterative proportional fitting algorithm can be used to fit the margins of the table after the perturbation according to the original margins. This results in cell values that are not integers. Exact additivity with integer counts can be achieved for simple tables by controlled rounding to base 1 using for example Tau-Argus (Salazar-Gonzalez et al. [33]). Cell values can also be rounded to their nearest integers resulting in “close” additivity because of the invariance property of the transition matrix. Finally, the use of microdata keys can also be adapted to this SDL method to ensure consistent perturbation of same cells across different tables by fixing the seed for the perturbation.

9.4 Differential Privacy in Survey Sampling and Perturbation

Shlomo and Skinner [36] explore how definitions of differential privacy, as introduced in the computer science literature [12, 15], relate to the disclosure risk scenarios of survey microdata. In addition, they investigate whether best practices at statistical agencies for the disclosure limitation of survey microdata would meet the strict privacy guarantees of differential privacy.

The disclosure risk scenario for survey microdata is that an intruder knows the values of some *key variables* for a target unit and seeks to use these values to link the unit to a record in the microdata after SDL has been applied, for n units in a sample s drawn from a population U . For identification risk to be well-defined, we assume that the records in the released microdata can meaningfully be associated with units in the population.

In the more recent computer science literature on privacy, there is usually no distinction between key variables and sensitive variables. The starting point is the

(original) database of attribute values from which the microdata are generated via the SDL method. It is supposed that an intruder wishes to learn about the attribute values for a specific (target) unit in the database. A “worst case” scenario is allowed for, in which the intruder has complete information about all other units represented in the database [15]. Under this assumption, let x denote the cell value, taking possible values $1, \dots, k$, where the contingency table is now formed by cross-classifying all variables, whether key or sensitive.

In the survey setting, there are two possible definitions of the database: the population “database” $\mathbf{x}_U = (x_1, \dots, x_N)$ and the sample “database” $\mathbf{x}_s = (x_1, \dots, x_n)$, where N denotes the size of the population $U = \{1, \dots, N\}$ and without loss of generality, we write $s = \{1, \dots, n\}$. The sample database might be viewed from one perspective as more realistic, since it contains the data collected by the statistical agency, whereas the population database would include values of survey variables for non-sampled units, which are unknown to the agency. In the context of differential privacy, we use the population database \mathbf{x}_U to define privacy and treat the sampling as part of the SDL mechanism, and suppose that prior intruder knowledge relates to aspects of \mathbf{x}_U .

Let \tilde{x}_i denote the cell value of unit i in the microdata after SDL has been applied, and let $\tilde{f}_j = \sum_{i \in s} I(\tilde{x}_i = j)$ denote the corresponding observed count in cell j in the microdata. Supposing that the SDM method leads to an arbitrary ordering of the records in the microdata, we can view the released data as the vector of counts: $\tilde{\mathbf{f}} = (\tilde{f}_1, \tilde{f}_2, \dots, \tilde{f}_k)$. Let $Pr[\tilde{\mathbf{f}} | \mathbf{x}_U]$ denote the probability of $\tilde{\mathbf{f}}$ with respect to an SDM mechanism, which includes sampling and/or misclassification, and where \mathbf{x}_U is treated as fixed. Shlomo and Skinner [36] consider the following definition.

Definition 9.1. ε -differential privacy [15] holds if:

$$\max \left| \ln \left(\frac{Pr[\tilde{\mathbf{f}} | \mathbf{x}_U^{(1)}]}{Pr[\tilde{\mathbf{f}} | \mathbf{x}_U^{(2)}]} \right) \right| \leq \varepsilon \quad (9.18)$$

for some $\varepsilon > 0$, where the maximum is over all pairs $(\mathbf{x}_U^{(1)}, \mathbf{x}_U^{(2)})$, which differ in only one element across all possible values of $\tilde{\mathbf{f}}$. The disclosure risk in this definition is inferential disclosure, where if only one value is changed in the population database the intruder is unable to gain any new knowledge about a target individual given that all other individuals in the population are known.

An (ε, δ) -probabilistic differential privacy definition, as discussed in [8], holds if (9.18) applies with probability at least $1 - \delta$ for some $\varepsilon, \delta > 0$. More precisely, this definition holds if the space of possible outcomes $\tilde{\mathbf{f}}$ may be partitioned into “good” and other outcomes, if (9.18) holds when the outcome is “good” and if the probability that the outcome is “good” is at least $1 - \delta$. This definition is essentially the same as the notion of probabilistic differential privacy in Machanavajjhala, et al. [25], where the set of “bad” outcomes is referred to as the *disclosure set*.

In the differential privacy literature, strict privacy guarantees are ensured by introducing random noise to the query. Laplace noise addition is proposed and the magnitude of the noise depends on the size of the query and parameter ε .

In Shlomo and Skinner [36], the aim is to investigate whether common practices at statistical agencies for preserving the confidentiality of respondents in survey microdata, such as sampling and perturbation, are differentially private. As an example from their paper, assume microdata are released from a sample which was drawn under an equal probability design. For this sample design and many other sampling schemes, f_j may equal F_j with positive probability. In particular, it is usual, in order to achieve unbiased statistical estimation, for an agency to require of any sampling scheme that all population units have positive inclusion probability and so, if any unit is population unique ($F_j = 1$) there is a positive probability that this unit is sampled, in which case $f_j = F_j = 1$. Thus, for a given \mathbf{f} and any sampling scheme where some element f_j of \mathbf{f} may equal F_j with positive probability, there exists a database $\mathbf{x}_U^{(1)}$ such that $f_j = F_j^{(1)} \geq 1$ for some j and $Pr[\mathbf{f}|\mathbf{x}_U^{(1)}] \neq 0$. Now, if we change an element of $\mathbf{x}_U^{(1)}$ which takes the value j to construct $\mathbf{x}_U^{(2)}$, for which $F_j^{(2)} = F_j^{(1)} - 1 < f_j$, we obtain $Pr[\mathbf{f}|\mathbf{x}_U^{(2)}] = 0$. Hence, ε -differential privacy does not hold for a very broad class of sampling schemes.

Shlomo and Skinner [36] present three reasons why the disclosure implications of this finding might not be considered a cause for concern by a statistical agency.

First, consider the threat that the event $f_j = F_j$ enables an intruder to disclose the cell value of a target individual. Such a disclosure depends upon the intruder knowing the count for the cell j across the whole of the population, excluding the target individual. Given this knowledge and the observation that this count equals $f_j - 1$, the intruder could infer that the target individual falls in this cell (and appears in the microdata). For the kinds of large populations of individuals upon which social surveys are typically based, it may be deemed unrealistic for an intruder to have precise information on all individuals in the population except one. The nearest realistic possibilities are that there exist an external database which either (a) via full population information, enables the population count F_j to be determined together with the identities of these F_j individuals, or (b) provides identities of an unknown subset of population individuals in the cell. In neither of these cases would exact disclosure occur. In (a), the key variable value for the target individual would already be known to the intruder. In (b), there would be residual uncertainty.

Second, consider the threat of identification, where an intruder knows both that a target individual belongs to cell j and that the individual is population unique, i.e., $F_j = 1$. In this case, the target individual is sampled (so that $f_j = 1$) then the intruder would be able to identify the individual in the microdata. This possibility is already well-known to agencies as a threat and grounds for ensuring that no microdata are released for which there are combinations of key variables for which an intruder could know that $F_j = 1$ or some other small value, such as in the kind of external database mentioned above.

Third, for any given database, the possible values of \mathbf{f} where ε -differential privacy fails may occur only with negligible probability. Therefore, the agency may consider it more appropriate to adopt the (ε, δ) -probabilistic differential privacy definition referred to earlier.

To explore the probabilistic nature of the threat to privacy further, consider the event that $f_j = F_j$, viewed here as the key threat to ε -differential privacy. By assumption, units in social surveys have small inclusion probabilities and the probability that all population units in a cell j will appear in the sample, i.e., $f_j = F_j = m$, will be very small for $m = 2$ (doubles) and even smaller for $m > 2$. The most realistic outcome is that a sample unique is population unique, i.e., the case $f_j = F_j = 1$, but this will typically also be possible with only a small probability. Therefore, while sampling does not guarantee differential privacy, it does provide probabilistic differential privacy under certain conditions.

Shlomo and Skinner [36] also examine perturbation under a misclassification mechanism:

$$Pr(\tilde{x}_i = j_1 | x_i = j_2) = M_{j_1 j_2}, \quad i = 1, \dots, n, \quad j_1, j_2 = 1, \dots, k \quad (9.19)$$

where \tilde{x}_i denotes the i -th element of $\tilde{\mathbf{x}}_s$. They found that ε -differential privacy holds if and only if all elements of \mathbf{M} are positive.

For example, under the SDL method of recoding, which is a common SDL technique for microdata arising from social surveys, assume a variable where categories 1 to α are grouped to category 1. The misclassification matrix is:

$$M_{j_1 j_2} = \begin{cases} 1, & \text{if } j_2 = 1, \dots, \alpha \text{ and } j_1 = 1, \text{ or} \\ & j_2 = \alpha + 1, \dots, k \text{ and } j_1 = j_2 - \alpha + 1 \\ 0, & \text{otherwise} \end{cases}$$

It is clear that with elements equal to zero, ε -differential privacy will not be guaranteed. Another example is the SDL technique of PRAM which uses a misclassification (probability) matrix \mathbf{R} to make random changes across categories of a variable as described in Sect. 9.2.2.1. The misclassification matrix should be defined to have no zero elements in order to ensure ε -differential privacy. Note that in practice, there may be zero elements in the misclassification matrix which represent structural zeros in the population, i.e. impossible combinations of categories such as children having an occupation as a “doctor”. For these cases, differential privacy is not applicable.

In general, Shlomo and Skinner [36] found that non-perturbative methods of SDL will not guarantee ε -differential privacy but may under some circumstances uphold (ε, δ) -probabilistic differential privacy, whereas perturbative methods which ensure no zero elements in the misclassification matrix will guarantee ε -differential privacy.

9.5 Future Outlook for Releasing Statistical Data

Statistical agencies are beginning to examine new dissemination strategies in order to allow more flexibility in the release of statistical data, whilst preserving the

confidentiality of respondents. More modern dissemination via web-based remote servers and table generators have now forced statistical agencies to consider stricter privacy guarantees offered by differential privacy.

As defined in Sect. 9.4, differential privacy aims to avoid inferential disclosure by ensuring that an intruder cannot make inference about a single unit when only one of its value is changed, given that all other units in the population are known. This definition would include disclosure by differencing and disclosure from highly predictive models, which are at the forefront of disclosure risk scenarios when considering flexible dissemination and online query systems compared to traditional outputs. The solution for guarantying differential privacy in the computer science literature is by adding noise/perturbation to the outputs of the queries under specific parameterizations.

In this section, we examine some new dissemination strategies that are being considered by statistical agencies.

9.5.1 Safe Data Enclaves and Remote Access

To meet increasing demands for high resolution data, many statistical agencies have set up data enclaves on their premises where approved researchers can go onsite and gain access to confidential statistical data. The secure servers within the enclave have no connection to printers or the internet and only authorized researchers are allowed to access them. To minimize disclosure risk, no data is to be removed from the enclave and researchers undergo specialized training to understand the confidentiality guidelines. Researchers are generally provided with standard software within the system, such as STATA, SAS and R, but any specialized software would not be available. All information flow is controlled and monitored. Any outputs to be taken out of the data enclave are dropped in a folder and manually checked by experienced confidentiality officers for disclosure risks. Examples of disclosure risks in outputs are small cell counts in tables, residual plots from regression models which may highlight outliers and Kernel density estimation with small band-widths.

The disadvantage of the data enclave is the need to travel, sometimes long distances, to access confidential data. In recent years, some statistical agencies have piloted remote access by extending the concept of the data enclave to a “virtual” data enclave. These “virtual” data enclaves can be set up at other government agencies, universities and even on a researcher’s own laptop. Users log on to secure servers via vpn connections to access the confidential data. All activity is logged and audited at the keystroke level and outputs are reviewed remotely by confidentiality officers before being sent back to the researchers via a secure file transfer protocol site.

9.5.2 Web-Based Applications

In recent years there are two types of web-based dissemination applications that are being considered by statistical agencies, namely *flexible table generators* and *remote analysis servers*.

9.5.2.1 Flexible Table Generating Servers

Driven by demand from policy makers and researchers for specialized and tailored tables from statistical data, particularly census data, some statistical agencies are developing flexible table generating servers that allow users to define and generate their own tables. The United States Census Bureau and the Australian Bureau of Statistics have developed such servers for disseminating census tables. Moreover, the Israel Central Bureau of Statistics developed a server for disseminating tables from the Social Survey. Users access the servers via the internet and define their own table of interest from a set of pre-defined variables and categories, typically provided in the form of drop down lists.

When selecting the SDL method to apply to the output table, there are two approaches: *apply SDL to the underlying data so that all tables generated in the server are deemed safe for dissemination* (known as *pre-tabular SDL*), or *produce tables directly from original data and apply the SDL method to the final tabular output* (known as *post-tabular SDL*). Although sometimes a neater and less resource intensive for data from a single source, the pre-tabular approach is problematic since it will compound the SDL impact and overprotect the data whilst increasing information loss. The post-tabular approach is also motivated by the computer science definition of differential privacy as discussed in Sect. 9.4.

For flexible table generation, the server has to quantify the disclosure risk in the original table, apply an SDL method on the data and then reassess the disclosure risk. Obviously, the disclosure risk will depend on whether the underlying data is a whole population (census) and the zeros are real zeros, or the data are from a survey and the zeros may be random zeros. After the table is protected, the server should also calculate the impact on information loss by comparing the perturbed table to the original data table.

Measures based on Information Theory were presented in Sect. 9.3.2 and can be used to assess disclosure risk and information loss in a table generating server since they can be calculated on-the-fly. In addition, some perturbation methods for protecting census tables were presented in Sect. 9.3.3.

The design of remote table generating servers typically involves many ad-hoc preliminary SDL rules that can easily be programmed within the system in order to determine whether tables can be safely released or not. These SDL rules may include limiting the number of dimensions in the table, minimum population thresholds, ensuring consistent and nested categories of variables to avoid disclosure by differencing, and so on.

Table 9.1 Disclosure risk and information loss for the generated table

	Disclosure risk	Hellinger distance
<i>Original</i>	0.318	–
<i>Perturbed input</i>		
Record swapping	0.282	0.988
Semi-controlled random rounding	0.137	0.991
Stochastic perturbation	0.239	0.995
<i>Perturbed output</i>		
Semi-controlled random rounding	0.135	0.993

As an example, in Table 9.1 we compare different SDL methods for a census table defined in one region of the United Kingdom according to banded age groups, education qualification and occupation. The table contained 2,457 cells where 62.4% were real zeros. The underlying data in the flexible table generating server was a very large hypercube which provides a priori protection since no units below the level of the cells of the hypercube are disseminated. We compare three pre-tabular methods on the hypercube: record swapping, semi-controlled random rounding and a stochastic perturbation, and a post-tabular method of semi-controlled random rounding applied directly to the output table. The measures are based on Information Theory, as described in Sect. 9.3.2.

From Table 9.1, it is clear that the method of record swapping when applied to the input data did little to reduce the disclosure risk in the final output table. This was due to the fact that the small cells remain unperturbed in the table. Record swapping provided the lowest data utility since the geography variable that was swapped was used to select a sub-population of the table thus increasing the information loss. From among the input perturbation methods, the semi-controlled random rounding provided the most protection against disclosure. The stochastic perturbation still leaves small cells in the table and hence is not as protective as the semi-controlled random rounding. Both methods have similar information loss. Comparing the pre-tabular and post-tabular semi-controlled random rounding procedure, we see slightly lower disclosure risk according to the post-tabular rounding and slightly higher data utility since the SDL method is not compounded by aggregating rounded cells.

9.5.2.2 Remote Analysis Servers

A remote analysis server is an online system which accepts a query from the researcher, runs it within a secure environment on the underlying data and returns a confidentialized output without the need for human intervention to manually check the outputs for disclosure risks. Similar to flexible table generators, the queries are submitted through a remote interface and researchers do not have direct access to

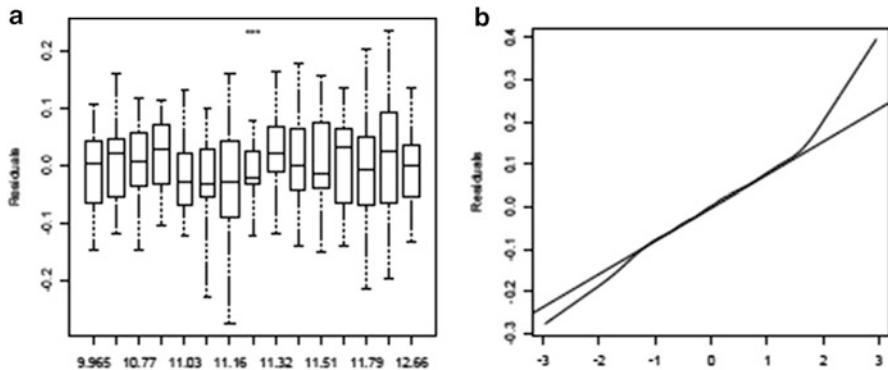


Fig. 9.1 Confidential residual plot from a regression analysis on receipts for the Sugar Canes dataset. (a) Residuals by fitted values. (b) Normal QQ plot of residuals

the data. The queries may include exploratory analysis, measures of association, regression models and statistical testing. The queries can be run on the original data or confidentialized data and may be restricted and audited depending on the level of required protection. O’Keefe and Good [27] describe regression modeling via a remote analysis server.

O’Keefe and Shlomo [26] compared outputs based on original data and two SDL approaches: outputs from confidentialized microdata and confidentialized outputs obtained from the original data via a remote analysis server. The comparison was carried out on a dataset from the 1982 survey of the sugar cane industry in Queensland, Australia (Chambers and Dunstan [7]). The dataset corresponds to a sample of 338 Queensland sugar farms and contained the following variables: region, area, harvest, receipts, costs and profits (equal to receipts minus costs). The dataset was confidentialized by deleting large outlier farms, coarsening the variable area and adding random noise to harvest, receipts, costs and profits.

Figure 9.1 shows what the residual plots would look like in a remote analysis server, where the response variable is receipts and the explanatory variables: region, area, harvests and costs. As can be seen, the scatterplot is presented as sequential box plots and the Normal QQ plot is smoothed. Figure 9.2 presents the comparison of the univariate analysis of receipts on the original dataset, the confidentialized input approach and the confidentialized output approach.

9.5.3 Synthetic Data

Basic confidential data is a fundamental product of virtually all statistical agency programs. These lead to the publication of public-use products, such as summary data, microdata from surveys, etc. Confidential data may also be used for internal use within data enclaves. In recent years, there has been a move to produce synthetic

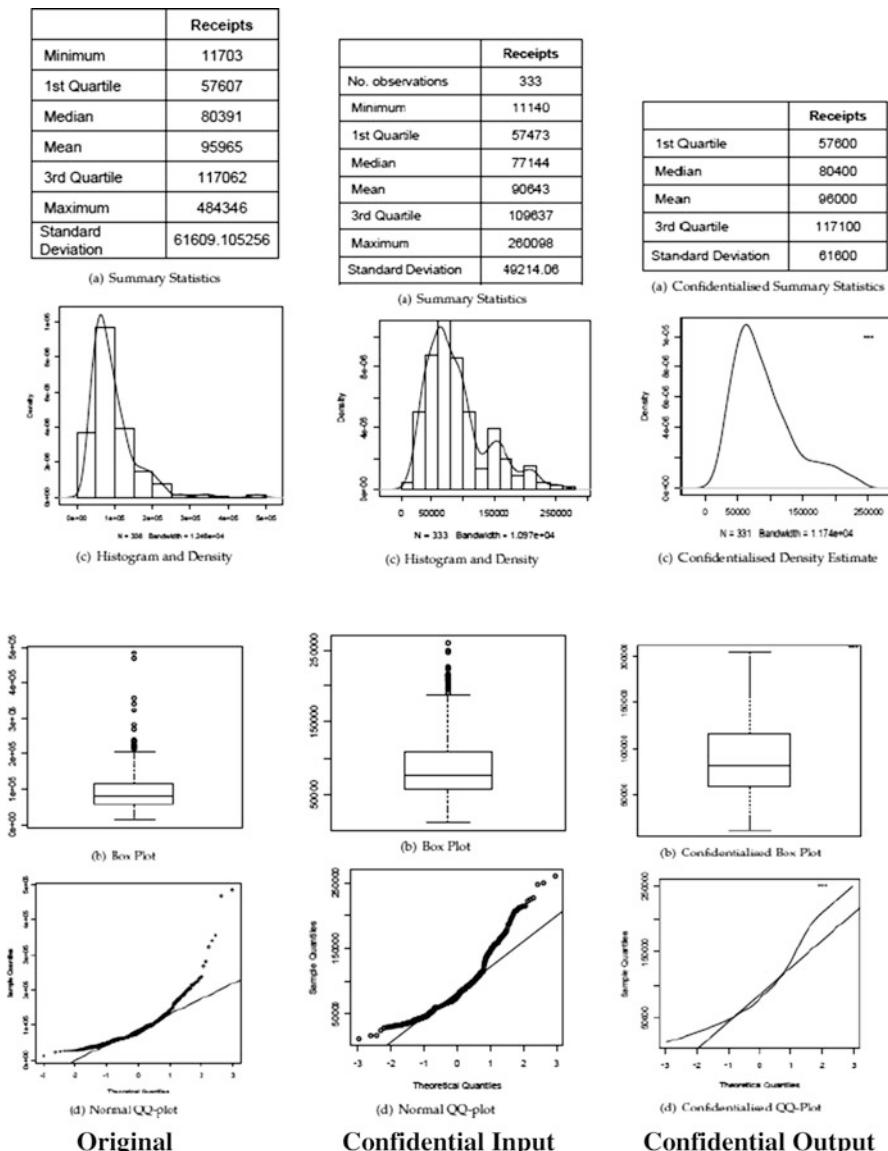


Fig. 9.2 Univariate analysis of receipts for the Sugar Canes dataset

microdata as public-use files which preserve some of the statistical properties of microdata. The data elements are replaced with synthetic values sampled from an appropriate probability model. The model is fit to the original data to produce synthetic populations through a posterior predictive distribution similar to the theory of multiple imputation. Several samples are drawn from the population to take

into account the uncertainty of the model and to obtain variance estimates. See [10, 28, 29] and references therein for more details of generating synthetic data.

The synthetic data can be implemented on parts of data, so that a mixture of real and synthetic data is released [24]. One application which uses partially synthetic data is the US Census Bureau “On the Map” available at: <http://onthemap.ces.census.gov/>. It is a web-based mapping and reporting application that shows where workers are employed and where they live according to the Origin-Destination Employment Statistics. More information is provided in Abowd and Vilhuber [1].

In practice it is very difficult to capture all conditional relationships between variables and within sub-populations. If models used in a statistical analysis are sub-models of the model used to generate data, then the analysis of multiple synthetic samples should give valid inferences. In addition, partially synthetic datasets may still have disclosure risks and need to be checked prior to dissemination.

9.6 Conclusion

In recent years, statistical agencies have been restricting access to statistical data due to their inability to cope with the large demand for high-resolution data whilst ensuring the confidentiality of statistical units. However, with government initiatives for ‘open data’, new ways to disseminate statistical data are being explored. This has led to more cooperation with computer scientists who have developed formal definitions of disclosure risk, particularly for inferential disclosure, and the adaption of SDL techniques that have stricter privacy guarantees. However, these techniques come at a cost in that researchers will have to cope with analyzing perturbed data which will require more training in the area of statistical inference under measurement (perturbation) errors. The methods for coping with measurement errors in statistical modelling requires that the statistical agencies release the parameters that are used in generating the SDL technique, for example, the variance of the additive noise and the swap rate in record swapping. Without these parameters, researchers will not be able to account for the measurement/perturbation errors in their analysis.

References

1. Abowd, J.M., Vilhuber, L.: How protective are synthetic data? In: J. Domingo-Ferrer, Y. Saygn (eds.) *Privacy in Statistical Databases*. Lecture Notes in Computer Science, vol. 5262, pp. 239–246. Springer, Heidelberg (2008)
2. Antal, L., Shlomo, N., Elliot, M.: In: J. Domingo-Ferrer (ed.) *Privacy in Statistical Databases*. Lecture Notes in Computer Science, vol. 8744, pp. 62–78. Springer International Publishing, New York (2014)
3. Anwar, N.: Micro-aggregation – the small aggregates method. *Informe Intern.* Eurostat, Luxembourg (1993)

4. Benedetti, R., Capobianchi, A., Franconi, L.: Individual risk of disclosure using sampling Design information. Istat: Contributi (2003) http://www3.istat.it/dati/pubbsci/contributi/Contributi/contr_2003/2003_14.pdf
5. Bethlehem, J., Keller, W., Pannekoek, J.: Disclosure limitation of microdata. *J. Am. Stat. Assoc.* **85**, 38–45 (1990)
6. Brand, R.: Microdata protection through noise addition. In: J. Domingo-Ferrer (ed.) *Inference Control in Statistical Databases*. Lecture Notes in Computer Science, vol. 2316, pp. 97–116. Springer, Heidelberg (2002)
7. Chambers, R.L., Dunstan, R.: Estimating distribution functions from survey data. *Biometrika* **73**(3), 597–604 (1986)
8. Chaudhuri, K., Mishra, N.: When random sampling preserves privacy. In: C. Dwork (ed.) *Advances in Cryptology - CRYPTO 2006*. Lecture Notes in Computer Science, vol. 4117, pp. 198–213. Springer, Berlin (2006)
9. Dalenius, T., Reiss, S.P.: Data swapping: a technique for disclosure limitation. *J. Stat. Plann. Inference* **7**(1), 73–85 (1982)
10. Dandekar, R.A., Cox, L.H.: Synthetic tabular data: an alternative to complementary cell suppression. Energy Information Administration, U.S. Department of Energy (2002)
11. Defays, D., Nanopoulos, P.: Panels of enterprises and confidentiality: the small aggregates method. In: *Proceedings of Statistics Canada Symposium 92, Design and Analysis of Longitudinal Surveys*, p. 195204 (1992)
12. Dinur, I., Nissim, K.: Revealing information while preserving privacy. In: *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '03)*, pp. 202–210. Association for Computing Machinery (2003)
13. Domingo-Ferrer, J., Mateo-Sanz, J., Torra, V.: Comparing sdc methods for micro-data on the basis of information loss and disclosure risk. In: *Proceedings of the ETK-NTTS Conference* (2001)
14. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. Knowl. Data Eng.* **14**(1), 189–201 (2002)
15. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: S. Halevi, T. Rabin (eds.) *Theory of Cryptography*. Lecture Notes in Computer Science, vol. 3876, pp. 265–284. Springer, Berlin (2006)
16. Elamir, E.A., Skinner, C.J.: Record level measures of disclosure risk for survey microdata. *J. Off. Stat.* **22**(3), 525–539 (2006)
17. Fienberg, S., McIntyre, J.: Data swapping: variations on a theme by dalenius and reiss. *J. Off. Stat.* **9**(1), 383–406 (2005)
18. Fraser, B., Wooton, J.: A proposed method for confidentialising tabular output to protect against differencing. In: *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality* (2005)
19. Fuller, W.A.: Masking procedures for micro-data disclosure limitation. *J. Off. Stat.* **9**(1), 383–406 (1993)
20. Gomatam, S., Karr, A.: Distortion measures for categorical data swapping. Technical Report Number 131, National Institute of Statistical Sciences (2003)
21. Gouweleeuw, J., Kooiman, P., Willenborg, L., De Wolf, P.: Post randomisation for statistical disclosure limitation: theory and implementation. *J. Off. Stat.* **14**(1), 463–478 (1998)
22. Hundepool, A.: The casc project. In: J. Domingo-Ferrer (ed.) *Inference Control in Statistical Databases*. Lecture Notes in Computer Science, vol. 2316, pp. 172–180. Springer, Berlin (2002)
23. Kim, J.: A method for limiting disclosure in micro-data based on random noise and transformation. In: *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 370–374 (1986)
24. Little, R., Liu, F.: Selective multiple imputation of keys for statistical disclosure control in microdata. The University of Michigan Department of Biostatistics Working Paper Series. Working Paper 6. (2003)

25. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: theory meets practice on the map. In: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering (ICDE '08), pp. 277–286. IEEE (2008)
26. O'Keefe, C.M., Shlomo, N.: Comparison of remote analysis with statistical disclosure control for protecting the confidentiality of business data. *Trans. Data Privacy* **5**(2), 403–432 (2012)
27. O'Keefe, C.M., Good, N.M.: A remote analysis server - what does regression output look like? In: J. Domingo-Ferrer, Y. Saygn (eds.) *Privacy in Statistical Databases*. Lecture Notes in Computer Science, vol. 5262, pp. 270–283. Springer, Berlin (2008)
28. Raghunathan, T.E., Reiter, J.P., Rubin, D.B.: Multiple imputation for statistical disclosure limitation. *J. Off. Stat.* **19**(1), 1–16 (2003)
29. Reiter, J.: Releasing multiply imputed, synthetic public-use microdata: an illustration and empirical study. *J. R. Stat. Soc. A* **168**(1), 185–205 (2005)
30. Rinott, Y., Shlomo, N.: A generalized negative binomial smoothing model for sample disclosure risk estimation. In: J. Domingo-Ferrer, L. Franconi (eds.) *Privacy in Statistical Databases*. Lecture Notes in Computer Science, vol. 4302, pp. 82–93. Springer, Berlin (2006)
31. Rinott, Y., Shlomo, N.: A smoothing model for sample disclosure risk estimation. In: *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*. Institute of Mathematical Statistics, Lecture Notes Monograph Series **54**, 161–171 (2007)
32. Rinott, Y., Shlomo, N.: Variances and confidence intervals for sample disclosure risk measures. Proceedings of the 56th World Statistics Conference Lisboa, Portugal, Instituto Nacional de Estatística (INE), 1090–1096 (2007) <http://isi.cbs.nl/iamamember/CD7-Lisboa2007/Bulletin-of-the-ISI-Volume-LXII-2007.pdf>
33. Salazar-Gonzalez, J.J., Bycroft, C., Stagemeier, A.T.: Controlled rounding implementation. In: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva, 9–11 Nov 2005
34. Shlomo, N., Antal, L., Elliot, M.: Measuring disclosure risk and data utility for flexible table generators. *J. Off. Stat.* **31**(2), 305–324 (2015)
35. Shlomo, N.: Statistical disclosure limitation methods for census frequency tables. *J. Int. Stat. Rev.* **75**(2), 199–217 (2007)
36. Shlomo, N., Skinner, C.: Privacy protection from sampling and perturbation in survey microdata. *J. Privacy Confidentiality* **4**(1), 155–169 (2012)
37. Shlomo, N., Skinner, C.: Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. *Ann. Appl. Stat.* **4**(3), 1291–1310 (2010)
38. Shlomo, N., De Waal, T.: Protection of micro-data subject to edit constraints against statistical disclosure. *J. Off. Stat.* **24**(2), 1–26 (2008)
39. Shlomo, N., Skinner, C.: Privacy protection from sampling and perturbation in survey microdata. *J. Privacy Confidentiality* **4**(1), 155–169 (2012)
40. Shlomo, N., Young, C.: Statistical disclosure control methods through a risk-utility framework. In: Proceedings of the 2006 CENEX-SDC Project International Conference on Privacy in Statistical Databases (PSD'06), pp. 68–81. Springer (2006)
41. Shlomo, N., Young, C.: Invariant post-tabular protection of census frequency counts. In: J. Domingo-Ferrer, Y. Saygn (eds.) *Privacy in Statistical Databases*. Lecture Notes in Computer Science, vol. 5262, pp. 77–89. Springer, Berlin (2008)
42. Skinner, C., Holmes, D.: Estimating the re-identification risk per record in microdata. *J. Off. Stat.* **14**(1), 361–372 (1998)
43. Skinner, C., Shlomo, N.: Assessing identification risk in survey micro-data using log-linear models. *J. Am. Stat. Assoc.* **103**(483), 989–1001 (2008)
44. Willenborg, L., De Waal, T.: Elements of statistical disclosure limitation in practice. In: *Lecture Notes in Statistics*, vol. 155. Springer, New York (2001)

Part II

**Privacy in Distributed and Dynamic
Settings**

Chapter 10

A Review of Privacy Preserving Mechanisms for Record Linkage

Luca Bonomi, Liyue Fan, and Li Xiong

Abstract Record linkage represents the process of identifying records that refer to the same real-world entity across multiple sources. In the recent past, the large explosion of data from organizations and individuals has created critical challenges in record linkage process, leading the scientific community to develop a rich series of techniques. Among these recent developments, the design of record linkage solutions for medical data is a cornerstone for health-care systems, as personal identifying information, such as name and date of birth, is often used for linkage. In this setting, record linkage solutions are expected to preserve individual patients' privacy in addition to be effective and efficient. In this chapter, we provide a broad review of the recent works in privacy preserving record linkage (PPRL). From a privacy-centric perspective, we summarize a comprehensive framework of the PPRL process and describe the privacy assurance techniques adopted for each step in the framework. With the applications in biomedical domain in mind, we identify several challenges for PPRL and discuss future research directions.

10.1 Introduction

Record linkage, also referred to as *duplicate detection* [16] or *entity resolution* [19], is a process that identifies records which belong to the same real-world entity across two or more data sources. Record linkage relies primarily on matching of names, addresses, and other fields that are typically not unique identifiers of entities. Given records in two datasets A and B , these can be represented as vectors in which each component is associated with a record's field or attribute. For example, in

This work was done while the author “Liyue Fan” was at Emory University.

L. Bonomi (✉) • L. Xiong
Department of Mathematics and Computer Science, Emory University, Atlanta, GA, USA
e-mail: lbonomi@emory.edu; lxiong@emory.edu

L. Fan
Information Laboratory (InfoLAB), University of Southern California, Los Angeles, CA, USA
e-mail: liyuefan@usc.edu

a medical setting, two hospitals interested in reconstructing their patients' history may represent each record in their databases as a vector, where the first component is associated with the patient's name, the second with her address, and the other components mapping the remaining attributes. Then, under this representation the records can be matched by examining the similarity between each component in the vectors.

In the original work proposed by Fellegi and Sunter [17], the record linkage approach is described as the process of partitioning the cartesian product of the datasets $A \times B$ into three disjoint sets: M the set of "matches", U the set of "non-matches", and C a set which requires human intervention in order to classify. The classification of the records into these sets is performed by using similarity functions that can differ on the particular field in the records (e.g., character-based similarity metrics to handle typographical errors). Therefore, the typical record linkage process consists first in pre-processing the data in order to compute the cartesian product $A \times B$ and successively in matching the records with the intent of minimizing the elements that are classified in the set C .

In the biomedical domain, linking records across multiple databases can provide a comprehensive view of a patient's medical history to increase safety and to aid decision making. Due to the decentralized nature of healthcare systems, data records that correspond to the same patient are often distributed across multiple disparate providers. For instance, it is common for hospitals to use a patient's Social Security Number (SSN) and/or demographics information for administrative purposes [26]. However, due to federal regulations and organizations' policies, the disclosure of such attributes is strictly prohibited. Therefore, without a universal patient identifier, the integration of medical databases largely relies on available common attributes, such as names, addresses, dates of birth, etc. Such integration process could be extremely challenging due to missing values or errors present in the record attributes. Below, we report some of the regulations for medical data records.

Legal, Regulatory, and Ethical Constraints Medical data records contain sensitive information about private patients, such as genome sequences, test results, treatments, etc. The collection and release activities are governed by rules, regulations, and legislative authorizations. For example, the European Union's Data Protection Directive 95/46/EC (the "EU Directive") protects the privacy of all personal data collected for, or about, EU citizens [49]. Data is considered personal, or "identifiable", if it enables anyone to link information to a specific person, even if the data holders cannot make that link. The Health Insurance Portability and Accountability Act of (HIPAA) Privacy Rule [51] in United States regulates the use and disclosure of individually identifiable information by health plans, providers, and other covered entities. Under HIPAA's "safe harbor" standard, 18 listed identifiers must be removed from data before it can be shared with an outside party, including names, social security numbers, biometric identifiers and other individually unique information. In addition, ethical issues of health-related data sharing must be considered, which are not specified in legislation and regulation.

Uses of data other than those for which they are specifically collected should be ethically justified and meet some minimal standard, even if law or regulation allows these data to be shared.

Challenges Extensive studies have been performed on privacy-preserving record linkage (PPRL). Three major challenges can be identified in the privacy-preserving record linkage process: *privacy*, *linkage quality* and *scalability*.

1. The *privacy* challenge during the PPRL process comes from two sources. First of all, record attribute values that are used in PPRL may be disclosed or reconstructed and patients can be re-identified as a result. A variety of de-identification techniques have been proposed to provide privacy guarantee on the data level, such as: *k*-anonymity [48], encryption/hashing [44], and differential privacy [14]. Secondly, record attribute values may be inferred by parties that are involved in PPRL protocols. In fact, a party could perform such privacy attacks either by following the protocol and trying to infer sensitive information, as a *honest but curious adversary* (HBC) [21], or by deviating from the protocol in a malicious way, as a *malicious adversary* [21]. Existing PPRL approaches typically assume HBC parties, since malicious adversaries can cheat in an arbitrary way, which makes it hard to achieve privacy guarantees.
2. *Linkage quality* measures the ability of classifying candidate records into matches and non-matches. In real world cases, high linkage quality is hard to achieve due to the presence of errors, missing, or outdated values in the data records. Therefore, approximate matching techniques in the record matching phase are often employed to mitigate data heterogeneity. Furthermore, the privacy assurance techniques in the PPRL process pose additional challenges on the linkage quality. For example, certain data transformation techniques are prone to false positives [43, 45]; limited data types and operations are allowed due to encryption techniques. The PPRL approaches introduce a trade-off between privacy guarantees and quality of linkage.
3. Today's databases contain an enormous amount of records, which poses tremendous challenges on the *scalability* of the PPRL protocols. In fact, the number of potential record pair comparisons grows with the product of the cardinalities of original databases. To overcome this challenge, a series of blocking/indexing techniques have been proposed with the intent of removing record pairs that are likely non-matches, while preserving those pairs that could potentially be matched. In PPRL protocols, these indexing techniques are often combined with cryptographic techniques, e.g., [33, 53], to alleviate the high encryption cost by matching only to a small number of candidate pairs.

Scope A few works [12, 22, 52] survey techniques for a specific procedure in privacy-preserving record linkage, e.g., private string comparison [12], and provide a brief overview of individual PPRL protocols [22, 52]. The scope of this chapter differs from the above works in several ways.

In this chapter, we review the most recent works on PPRL, many of which have been specifically designed for biomedical applications. For instance, in Atallah et al.

[3] a secure protocol for computing *edit distance* is proposed, which can be used to measure the similarity between DNA sequences. The work of Du et al. [10] also investigates the problem of secure sequence matching and provides solutions for computing a few generic distance metrics.

This chapter provides a privacy-centric view of the PPRL process. We summarize a conceptual framework for PPRL, which includes *data transformation*, *blocking*, and *matching*, and review all privacy assurance techniques that have been proposed for each step by the existing research works. For example, Bloom Filter [45] and Lipschitz Embedding [43] can be used in the *data transformation* step to achieve data de-identification; Homomorphic encryption [33] and Secure Scalar Product [41] can be used in the *matching* step to measure pair-wise record similarity.

Through the framework, we categorize existing PPRL works according to the specific privacy techniques. We further provide a taxonomy with respect to *privacy*, *scalability*, and *linkage quality* to facilitate the understanding and implementation of these protocols. Toward the end, we point out several future research questions in PPRL, specifically related to applications to medical databases.

The organization of this chapter is as follows: Sect. 10.2 describes the PPRL model and provides a taxonomy of the existing privacy-preserving record linkage protocols. In Sect. 10.3, we review a set of secure data transformation techniques, and in Sect. 10.4 we summarize secure multi-party computation (SMC) protocols that are employed by PPRL approaches. Section 10.5 provides a set of blocking techniques, which can be used to reduce the number of pair-wise comparisons in an effort to improve the scalability of the PPRL process. Then, in Sect. 10.6 we describe some open research problems in applying privacy-preserving record linkage on medical datasets. Last, Sect. 10.7 concludes this chapter.

10.2 Overview of Privacy Preserving Record Linkage

This section provides an overview of the privacy-preserving record linkage process. Specifically, we illustrate the PPRL model and describe the major steps in the process. Aligned with this model, we provide a taxonomy of the approaches considered in the rest of the chapter and outline their key features.

10.2.1 The PPRL Model

The privacy preserving record linkage techniques typically include two data sources following a multi-step process, as illustrated in Fig. 10.1. We use this representation to highlight the steps where privacy preserving techniques are employed. In particular, we identify three major steps: *data transformation*, *blocking/indexing* and *record matching*, where the first two steps are optional. In fact, theoretically it is sufficient to achieve privacy by applying secure multi-party computation (SMC)

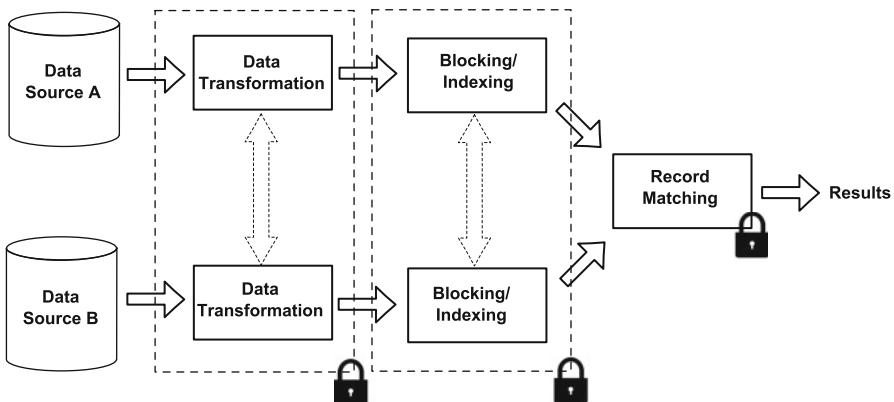


Fig. 10.1 The privacy-preserving record linkage (PPRL) model

protocols at the record matching step. However, since SMC protocols largely rely on cryptographic mechanisms, this straightforward solution could result in prohibitive computational cost for real-world applications.

Many works have proposed to employ data transformation and blocking/indexing techniques to achieve privacy, as well as to improve the computational efficiency. In the data transformation step, privacy is typically achieved using transformation schemes/anonymization techniques, so that the transformed records can be used in the remaining steps without incurring additional privacy manipulations. Although blocking/indexing methods have been historically designed to accelerate the linkage process, they can also be modified to preserve private information about the data records, when the data transformation step is not performed. Detailed descriptions of the framework are provided below.

Data Transformation Step Many PPRL protocols proposed in the literature perform a data transformation step to achieve privacy in the record linkage process. This step aims to generate new representations for the original records, which are then used in the rest of the record linkage process. The advantages of this phase are two-fold. Firstly, the transformation generally provides privacy by either using mapping functions that are hard to reverse or removing sensitive information from the original records. Secondly, the similarity properties of the original records can be typically preserved to some extent by the transformation, which enables record linkage techniques to match transformed records either in an exact or approximate way. As is shown in the framework, data owners may share some information during this step to improve the utility of transformed data; however, the transformation of data records is often done by the owner of the data.

Blocking/Indexing Step This phase is generally used to reduce the computational cost of record linkage. In fact, the blocking/indexing step aims to reduce the number of pair-wise record comparisons in the matching step, by removing those pairs that are unlikely to be linked. As a result, only the record pairs that represent potential

matches are further evaluated in the matching step. The blocking/indexing step does not typically introduce privacy guarantees in the PPRL process as a standard alone component, but rather it is used in combination with private data transformations or SMC protocols in the matching step. Similar to the transformation step, the blocking phase is generally performed by each data owner and in some cases information may be shared among parties to improve the quality of blocking.

Records Matching Step In this phase, the candidate record pairs are compared in detail to determine whether they represent a match. The pair-wise comparison between records can be performed either at *record level* or *attribute level*. In the first case, the attribute values are concatenated to form a long string, which is used in the similarity evaluation. On the other hand, in the second case the comparison is performed using specialized similarity functions between individual attribute values. Depending on the nature of the attribute type, different similarity measures can be employed for the comparison. For example, the edit distance [34] (i.e., Levenshtein distance) is commonly used to measure the similarity between string records. The comparison between records can be performed in an *exact* or in an *approximate* way. While in the former case, records are considered as a match when their corresponding values are exactly the same; in the latter, records are matched if their similarity score is above a certain threshold. The matching step can be either performed by a third party or by the data owners. However, in both cases it requires the data owners to share the candidate records. Therefore, SMC protocols are often employed in literature to achieve privacy in this step.

In this framework, the privacy of data records can be preserved by either converting them to new representations with *data transformation* techniques, or matching them securely through SMC protocols at the *matching* step. The *blocking/indexing* step, often deployed together with *data transformation* or *secure matching*, does not offer privacy protection alone in general. However, a few blocking methods have been proposed [26, 33] to provide k -anonymity and differential privacy guarantees, in case of knowledgeable adversaries.

10.2.2 Taxonomy of Presented Techniques

We provide a taxonomy for the privacy-preserving record linkage approaches presented in the rest of the chapter. Our goal is to provide a clear overview of the current techniques and to outline the differences in the privacy-preserving mechanisms employed by these approaches. A recent survey that presents a taxonomy of PPRL techniques along 15 dimensions can be found in Vatsalan et al. [52].

We distinguish the PPRL solutions into three major families: *secure transformations*, *secure multi-party computation (SMC)* and *hybrid* techniques, according to where the privacy notion is applied in the record linkage model. For each family, we provide a comprehensive view of the techniques, examining seven dimensions which we group into three categories: *privacy guarantee*, *scalability* and *linkage*

quality. The complete list of the PPRL approaches is shown in Table 10.1. In the rest of this section, we provide a brief overview of the features of the privacy-preserving record linkage protocols that are considered in this chapter.

10.2.2.1 Privacy Guarantee

A series of privacy methodologies have been developed to provide privacy guarantees in the record linkage process. The privacy notion employed by the solution is often related to the number of parties involved and the type of the protocol that is used. Typically, PPRL solutions involve two or three parties in the process. In the first case, only the data owners participate in the process and therefore solutions following this scheme are called *two-party* protocols. In the second case, a third party is present in the protocol with the task of matching record pairs. Solutions following this paradigm are called *three-party* protocols.

Privacy-preserving techniques, in the context of medical data, center around the following privacy notions: k -anonymity, encryption/hashing, and differential privacy. Here we briefly introduce these privacy notions.

k -Anonymity The k -anonymity notion [48] has been used in the privacy-preserving record linkage setting to construct a variety of privacy-preserving solutions [26, 27, 36]. Given an integer value k , a dataset satisfies k -anonymity if for each combination of quasi-identifiers (i.e., attributes that can be used to identify individual entities) there are at least k records in the dataset matching such configuration. Intuitively, for each individual record, there are other $k - 1$ records having the same values for the quasi-identifiers. In general, this privacy notion is used in the data transformation step as well as in the indexing/blocking step. To achieve k -anonymity, the attributes of the records in the original dataset are *suppressed* or *generalized*. While in the first case the attribute values are replaced with “**”, in the second case they are replaced with a broader range or category. An example database that satisfies the k -anonymity principle is illustrated in Table 10.2.

Encryption/Hashing Encryption and hashing techniques are widely used in the PPRL approaches. Typically, in these works, the data holders encrypt the identifiers with standard cryptographic procedures and send the encrypted data to the third party to perform the match. The data holders agree on the secret key to use in encrypting the identifiers in their databases and send only the encrypted data to a third party. Since the identifiers agree exactly if their corresponding hash values are the same, the third party can link matching records without knowing the identifiers.

Works in [43, 45], for example, perform data transformation to achieve privacy using Bloom Filter [5] with one-way hashing functions and embedding techniques, respectively. While approaches based on hashing functions match records by hashing the original data and computing similarity between the hashed results, embedding techniques instead map the original records in a vector space where pair-wise distance is generally preserved up to a distortion factor. Intuitively, the

Table 10.1 Overview of the PPRL techniques reviewed in the chapter.

PPRL Technique	Privacy Guarantee		Scalability		Linkage Quality		Data type
	Privacy Notion	# Parties	Indexing	Comp.	Matching	Comparison	
Mohammed et al. [36]	Attr. Generalization	2+	None	Polynomial-Log	Attribute	Approx.	Categorical
	Secure Hashing	3	None	Quadratic	Record	Approx.	String
Schnell et al. [45]	Bloom Filter						
	Secure Hashing	3	None	Quadratic	Record	Approx.	String
Durham et al. [13]	Bloom Filter						
	Secure Hashing	3	None	Quadratic	Attribute	Approx.	String
Churches and Christen [9]	Secure Hashing	3	Standard	Quadratic	Attribute	Approx.	String
	Secure Hashing	3	SMC	Quadratic	Attribute	Approx.	String
Al-Lawai et al. [2]	Embedding	3	None	Quadratic	Attribute	Approx.	String
	Embedding	2	Sorted	Quadratic	Attribute	Approx.	String
Scannapieco et al. [43]	SMC						
	Yakout et al. [53]						
Bonomi et al. [6]	Embedding	3	None	Quadratic	Attribute	Approx.	String
	Differential Privacy						
Karakasidis et al. [29]	Encoding	3	Clustering	Quadratic	Attribute	Approx.	String
	Differential Privacy	2	Standard	Quadratic	Attribute	Approx.	Categorical
Inan et al. [24]	SMC						

(continued)

Table 10.1 (continued)

PPRL technique	Privacy guarantee	Scalability	Linkage quality		
	# Parties	Indexing	Comp.	Matching	Comparison
Kuzu et al. [33]	Differential Privacy	3	Standard	Quadratic	Approx.
	SMC				
Karakasidis and Verykios [27]	Encoding	2	Mapping	Quadratic	Attribute
	SMC				
Du and Atallah [10]	SMC	3	None	Quadratic	Attribute
Li et al. [35]	SMC	2	None	Quadratic	Record
O'Keefe et al. [38]	SMC	3	None	Quadratic	Attribute
Atallah et al. [3]	SMC	2	None	Quadratic	Attribute
Kantarcioglu et al. [26]	k -anonymity	4	Standard	Quadratic	Record
	SMC				
Ravikumar et al. [41]	SMC	2	None	Quadratic	Record
Durham [11]	Secure Hashing Bloom Filter	3	Mapping	Quadratic	Record
Karakasidis and Verykios [28]	Encoding	3	Clustering	Quadratic	Attribute
	k -anonymity				
Pang et al. [40]	Encoding	3	Clustering	Quadratic	Attribute

Table 10.2 Example of k -anonymity, with $k = 2$

FIRST NAME	LAST NAME	AGE	ZIP CODE		First name	Last name	Age	Zip code
Johan	Smith	45	3203	→	*	Smith	[40–50]	32*
Susan	Stone	30	2692		Susan	*	[20–30]	26*
Beatrice	Smith	40	3230		*	Smith	[40–50]	32*
Susan	Delgrado	28	2609		Susan	*	[20–30]	26*

First Name and Last Name are suppressed, while Age and Zip Code are generalized

privacy guarantee that is provided by using hashing and embedding techniques is to transform the original records, hence reducing the chance of data re-identification or reconstruction by adversaries.

Approaches such as [2, 53] adopt SMC protocols in the record matching phase, where the candidate records are compared with cryptographic solutions. In general, SMC-based approaches allow the two party holders to directly perform the linkage of the records by following a communication protocol and encryption steps without requiring the presence of a third party. SMC solutions guarantee that at the end of the computation no party learns more than the final output (i.e., the final record linkage result). Despite the strong privacy notion, SMC protocols are computationally expensive due to the encryption schemes they employ. PPRL solutions based on SMC are generally designed with commutative encryption, homomorphic encryption, and secure scalar product protocols.

Differential Privacy Recently, Dwork [14] proposed a new privacy notion which aims to protect the individual records against adversaries with arbitrary background knowledge, when answering statistical queries on the database. The idea of differential privacy consists of perturbing the output of the query with random noise, such that an adversary cannot determine the presence or absence of any individual record. Differential privacy enables the protection of records without constraining the adversary model, which makes this approach very popular in the database field [15]. In the PPRL model, differential privacy can be applied when statistical queries are performed on the original datasets and the results of such queries are shared among parties. An example is the work of Inan et al. [24], where the notion of differential privacy is adopted to protect the block sizes in the indexing phase.

10.2.2.2 Scalability

The scalability for PPRL approaches is critical to the application of the proposed solutions in the real world. We examine each approach regarding the computation complexity and the indexing technique employed.

Complexity In a generic record linkage process, the number of record pair comparisons grows linearly with the products of the size of the two original databases and, therefore, a quadratic time complexity is expected. However, for PPRL approaches the privacy mechanisms could have substantial impact on the overall computational cost. For example, cryptographic schemes for pair-wise record matching could scale up the overall running time. In Table 10.1, we measure the complexity of privacy-preserving record linkage approaches with respect to the size of each original dataset, even though some approaches were originally proposed to solve one-vs-many matching problems, such as [10].

Indexing A widely used technique to reduce the number of pair-wise comparisons is indexing. It is achieved by grouping records into blocks and comparing only those that fall within the same block. In PPRL literature, a variety of indexing

techniques have been proposed that can be categorized into *standard blocking*, *sorted neighborhood*, *mapping* and *clustering*. When combined with private data transformation, these techniques can be performed on the transformed data records as well.

10.2.2.3 Linkage Quality

Linkage quality measures the ability of the PPRL techniques to determine matching and non-matching records. For each solution, we examine the type of comparison performed (record-level vs. attribute-level), the matching rule (exact vs. approximate matching), as well as the data type supported by the protocol. Clearly, those features are critical to the choice of the privacy mechanisms in PPRL approaches. For example, homomorphic encryption and scalar product protocols should be adopted to enable approximate matching as in [35], while commutative encryption can be adopted for exact matching, as in [41].

10.3 Secure Transformations

Secure transformation techniques aim to perform the linkage of the records after some transformations have been applied to the original data (typically strings), so that the privacy and security of the linkage is achieved at the first step of the record linkage process. The typical scenario involves three parties: two data owners and a third party in charge of the matching.

First, the data owners use secure transformation techniques to generate a new representation of the records and then send the transformed records to the third party which performs the matching. In this framework, many strategies have been proposed to achieve privacy in the transformation phase, which we distinguish in the following categories: *attribute suppression/generalization*, *n-grams methods*, *embedding*, and *phonetic encoding*. While attribute suppression/generalization techniques aim to achieve privacy by either removing or generalizing the sensitive information from the original records, the other three techniques generate a new representation for the record. It is worth noting that attribute removal techniques can be applied to generic data while *n*-grams, embedding and encoding are generally focused on string records. The main idea in the *n*-gram methods consists in breaking the string records into substrings of length *n* (*n*-grams) and use this information to represent the original data in the rest of the record linkage process. Embedding techniques map the original records in a vector space where the similarity distance between the records is typically preserved up to some distortion factors. Therefore, privacy-preserving record linkage approaches based on embedding techniques are able to perform both exact and approximate matching of the records. Encoding methods are generally used to generate a representation of the original strings which could be based on phonetic encoding, e.g., soundex.

In the rest of the section, we provide an overview of the secure transformation techniques and we describe the most relevant PPRL approaches presented in literature that fall in this category.

10.3.1 Attribute Suppression and Generalization Methods

A possible way to achieve privacy in the record linkage process is by removing the sensitive information from the patient's data records. In the medical domain, for example, the institutions involved in the record linkage process agree on exchanging de-identified information according to the HIPAA Privacy Rule [51]. Determining what information can be released without inappropriately compromising the privacy of the individual users is a crucial problem. In fact, the large amount of information available nowadays creates opportunities for adversaries to integrate the de-identified information with publicly available data with the intent of re-identifying the patient information. For example, Sweeney [47] showed that it was possible to re-identify the names and addresses of 97 % of the registered voters in Cambridge, Massachusetts, using the birth date and full postal code.

On the other hand, a stringent de-identification process could easily lead to unusable data for many research tasks in a variety of settings. For example, information such as geographical locations or treatment effects, when removed to prevent re-identification may lead to useless data for research studies such as epidemiological studies and pharmaceutical drug design.

In the recent years, researchers have focused on designing generalization techniques to reduce the possibility of re-identification but at the same time preserve the utility of the data. Among these works, the k -anonymity notion has been employed to achieve privacy by grouping each single record with other $k - 1$ records having the same attribute values. Privacy-preserving record linkage techniques typically employ k -anonymity either in the transformation or in the blocking phase. Below, we describe the PPRL approach recently proposed by Mohammed et al. [36], which uses this privacy notion by generalizing the attribute values.

Mohammed et al. [36] proposed a multi-party integration approach that achieves k -anonymity. The proposed technique is based on the framework originally developed by Jiang and Clifton [25] (DkA) to achieve distributed k -anonymity. Mohammed et al. extended the DkA framework and proposed two solutions for privately integrating records from multiple parties under different adversarial models. In the first approach, the authors consider a semi-honest adversarial model, which assumes that parties follow the protocol but may try to deduce additional information. In the second approach, the authors consider the presence of malicious parties that for their own benefit could deviate from the protocol. The anonymization of the data is achieved by employing a top-down generalization approach, where the values starting from the most general level are further specialized. The specialization of the attribute is decided according to the privacy level and the amount information disclosed to the other party. To prevent a malicious party to gain information by

deviating from the protocol in deciding the specialization level, the authors in [36] introduced game theory concepts to penalize those malicious parties that drastically deviate from the protocol.

10.3.2 N-Grams Methods

Privacy-preserving transformations based on the decomposition of the original string records in n -grams are extensively used by many approaches in literature. Such techniques are generally applied with the intent of achieving privacy, while at the same time providing high utility and low computational complexity. In fact, the decomposition of the original records into grams can be performed in an efficient manner and such representation preserves partial information about the strings that can be used to approximately match the records. The privacy for these techniques is generally achieved by using hash functions to encode the grams representation of the original string records.

The set of n -grams that is produced by breaking down the records can be either used to directly generate a vector representation of the original strings, or for hashing the grams into a Bloom filter [5]. The Bloom filter generates a bit array representing the original string, where each gram determines a position in the array according to a set of hash functions. Initially, all the bits are set to 0, and only the bits representing the hashed grams of the original string are set to 1. The final similarity

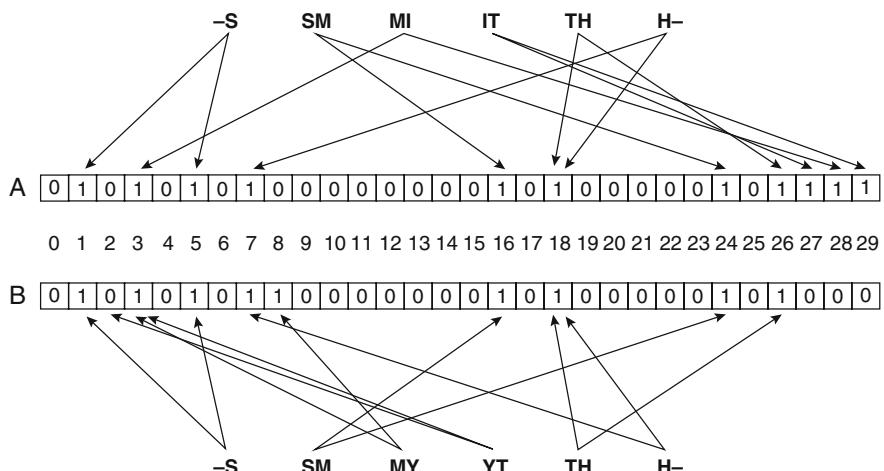


Fig. 10.2 Bloom Filter representation for the names SMITH and SMYTH using 2-g. The map is obtained using one hash function and in total there are ten bits in A, and 11 bits in B set to 1. Only eight bits are shared among the Bloom filters, therefore the similarity measure between the original strings approximated with the Dice coefficient is $\frac{2 \cdot 8}{(10+11)} \approx 0.762$ (example from Schnell et al. [45])

measure between two Bloom filter representations of the original strings is generally computed using set-based metrics, such as the Dice coefficient. An example of this representation is illustrated in Fig. 10.2.

In what follows, we describe some of the most recent privacy-preserving record linkage techniques that use n -grams based secure transformations.

Schnell et al. [45] proposed a secure transformation technique based on n -grams and Bloom Filters. Their approach is based on a three-party protocol, where the data owners transform their original data and the third party is in charge of the matching. The secure transformation is performed as follows. Each data owner concatenates the record attributes forming a single string, which is further decomposed in grams of length q . Then, the records are mapped into Bloom filters using a multitude of independent hash functions. In the original paper, the authors used the SHA1 and MD5 [44] hash functions to construct a Bloom filter with k hash functions without any increase in the asymptotic false positive probability. Their approach follows the idea of the double hashing schema proposed in [30]. After this transformation, the records are matched by the third party by computing the Dice coefficient on the Bloom filter representation of the original string records.

The representation of records via Bloom filters as in [45] could still disclose some information to an adversary. In fact, Kuzu et al. [32] proposed a constraint satisfaction cryptanalysis attack which exposes the vulnerability of the record representation based on Bloom filters. To improve the privacy, privacy-preserving record linkage techniques based on composite Bloom filters have been proposed in [13, 46]. The idea in these techniques is to first generate Bloom filters for single attributes (FBFs), then combine them into one composite Bloom filter per record (RBF). The construction of such a representation is illustrated in Fig. 10.3.

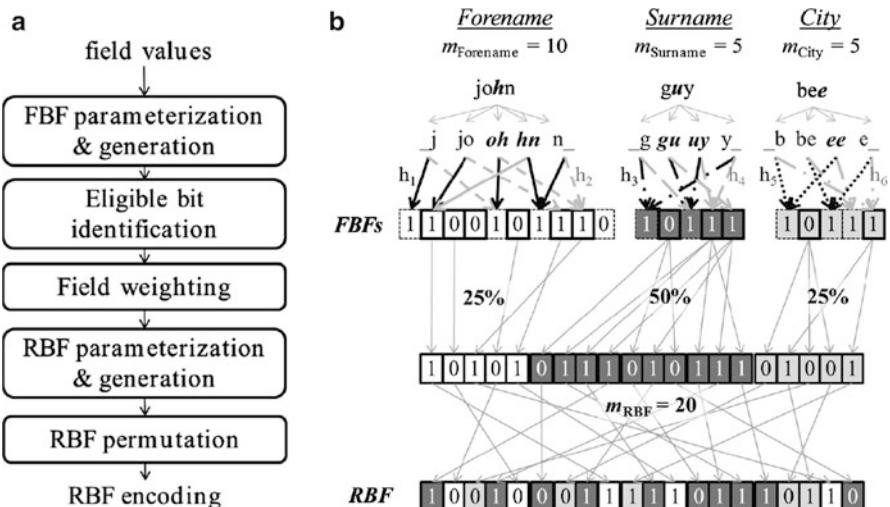


Fig. 10.3 Example of composite Bloom filter representation from Durham et al. [13]. (a) Transformation process. (b) Composite bloom filter

To increase the utility and the privacy for each attribute the number of bits selected and their weight may vary. The approach proposed by Durham et al. [13], which uses a record level Bloom filter representation, showed improved security and utility.

Churches and Christen [9] proposed a three party privacy-preserving record linkage protocol based on hashed n -grams. The data owners agree on a common hashing function which will be used to securely encrypt the data. Each original record is decomposed into grams and the power set of grams is computed by each data owner. Then, each set is hashed and sent to the third party which performs a secure equi-join on the transformed records. The use of subsets of grams enable the parties to perform approximate matching between their records by computing the Dice coefficient on the set of grams. However, due to the large number of grams in the power set, this approach is computationally intense. Furthermore, it has been shown in [50] that this approach is vulnerable to frequency attacks.

Al-Lawati et al. [2] employed a secure hashing technique based on TFIDF [42] (Token Frequency/ Inverse Document Frequency). Given a common vocabulary of m tokens among the two datasets, a set of m weights is computed for each original record by using the token frequency in the record (TF) and the token frequency in the whole dataset (IDF). This representation enables approximate matching between the original records, however it discloses some information about the original string. To achieve privacy, the authors in [2] proposed a secure hash transformation. A hash function H is evaluated on each token and it determines the index where the corresponding weight is stored in the new record representation. In a successive phase, the records are blocked and finally matched using a SMC protocol.

10.3.3 Embedding Methods

Embedding methods provide privacy by transforming the original data into a vector space, where the distance between the original records is preserved up to a distortion factor. This feature enables techniques based on embedding to determine the optimal threshold value in the new space to match records either in a exact or approximate way. Embedding-based approaches require a reference set (i.e., base for the embedding) to be used during the process of mapping the strings into a vector space. This set determines how the records are transformed and it can be either publicly available or privately held. In the first case, the reference set is known to all the parties involved in the protocol, while in the second scenario the reference set is kept secret from the third party. In this case, a higher level of security is achieved compared to the publicly available set. Some of the popular embedding approaches applied in the PPRL setting are described below.

Scannapieco et al. [43] proposed a three party protocol where the base for the embedding is formed by sets of random strings and it is privately determined by the data owners. This approach is based on the use of Sparse Map [23], which is a variant of the Lipschitz embedding. The key idea is to construct an embedding space using a set of reference sets, each containing strings randomly generated.

	Reference Set	SMITH	SMYT
S_1	BpuVH wyZrT	(4,4)	(5,4)
S_2	hWHJD aeplv	(4,5)	(5,5)
S_3	Zuawd QLUHo Zuawd QLUHo	(5,4,5,4)	(5,5,5,5)
S_4	wyZrT wyZrT fOOiw fOOiw	(4,4,5,5)	(4,4,5,5)

Embedding of SMITH, $v_1 = (4,4,4,4)$
 Embedding of SMYT, $v_2 = (4,5,5,4)$

Fig. 10.4 Embedding example for the names SMITH and SMYT with an embedding base formed by the sets S_1, S_2, S_3, S_4 of randomly generated strings

Each data owner embeds its own record, creating for each original string a vector representing the minimum distance (edit distance) between the original string and the strings in the base. In fact, the resulting vector is a vector of distances, where the i -th coordinate denotes the minimum distance of the original string from the i -th set in the base. The transformed records are subsequently sent to the third party, which determines the matching pairs by computing the Euclidean distance between vectors. An example of the embedding is illustrated in Fig. 10.4.

Yakout et al. [53] developed a two-party PPRL approach based on the idea proposed in [43]. In a first phase, the string records are mapped into vectors following the steps in [43] and for each vector a complex number is computed. In this complex plane, adjustable slabs are used to determine the likely matching pairs. Finally, in the second phase, a SMC protocol for secure scalar product is employed to compute the actual matching pairs of records.

Recently, Bonomi et al. [6] proposed a privacy-preserving record linkage technique, where the reference set for the embedding is computed in a privacy-preserving way directly from the original data. The idea is to use a data-dependent transformation to better capture the structure of the original data. Each data owner mines a set of frequent variable-length grams that are obtained using a prefix tree structure from the original records. The frequencies of the grams in the reference set satisfy the notion of differential privacy, which guarantees that an adversary cannot infer the presence of any individual string record in the original dataset by observing the frequency of the mined grams. After this operation, the data owners share the mined set and decide a common base for the embedding. Given the reference set with m grams, each original string record is embedded into a vector in \mathbb{R}^m , where the i -th coordinate represents the number of occurrences of the i -th gram in the original string scaled by the gram length. The transformed records are then sent to a third party, which performs approximate matching using the Euclidean distance.

Table 10.3 The Soundex encoding

Original symbols	Soundex value
a, e, h, i, o, u, w, y	0
b, f, p, v	1
c, g, j, k, q, s, x, z	2
d, t	3
l	4
m, n	5
r	6

10.3.4 Phonetic Encoding Methods

Phonetic encoding methods are used to generate a representation of the original string generally based on the phonetic properties, rather than the syntactic structure of the strings (e.g., grams decomposition). Soundex [7] is probably the best known phonetic encoding scheme that is used nowadays. The encoding is generated by keeping the first symbol in the original string and converting the rest into numbers. The mapping rule from symbol to number is reported in Table 10.3. All the zeros in the encoding are removed and sequences with occurrences of the same number are merged into one. The final code has fixed length of four symbols (one letter end three numbers), where short codes are padded with zeros and longer codes are stripped-off. This representation tends to be more robust than the n -gram method with respect to typos and use of nicknames in the strings, since this encoding is merely based on the phonetic properties of the string. For example, the names “John” and “Jon” are mapped to the same phonetic encoding “J500”.

A variety of PPRL techniques have been developed using phonetic encoding. For these methods the privacy is generally not provided by the encoding itself but rather it is achieved by hashing the encoded values using secure hash functions.

Karakasidis et al. [29] proposed three PPRL methods based on the Soundex phonetic encoding [7]. To reduce the vulnerability to frequency attacks, the authors introduce fake records which, together with the real records, are encoded and sent to the third party who performs the matching.

The first technique proposed is the Uniform Ciphertext/Uniform Plaintext (UCUP) approach, which aims to force the set of Soundex values and original string records to exhibit uniform distributions. Therefore, a number of fake records is injected, in a way that all the Soundex values map to an equal number of string records. Although this technique is intuitive and sound, it incurs a large complexity due to the huge number of fake records introduced to the original dataset.

In order to avoid oversized datasets, the authors in [29] introduced the Uniform Ciphertexts by Swapping Plaintexts (UCSP) strategy. In this method, the authors first compute the average number K of string records for each Soundex value, then remove the redundant occurrences of the string records with values larger than K and add an equal number of fake occurrences for Soundex values with fewer appearances. Although in this way no excessive fake records are introduced, those

records that have been removed will not participate to the matching phase and, therefore, a separate linking process should be initialized.

The last method introduced by the authors in [29] is the k -anonymous Ciphertexts approach (kaC). The method aims to create a dataset where each Soundex value is associated with at least k original string records. Therefore the parameter k provides a trade-off between the privacy and the efficiency of this approach.

10.4 Secure Multi-Party Computation

In order to compute the similarity of record pairs without disclosing private records, secure multi-party computation (SMC) approaches have been widely adopted by privacy-preserving record linkage protocols to identify the matching record pairs accurately and securely. Based on the core technique of secure computation, we observe that existing works fall into three categories: *commutative encryption* based protocols, *homomorphic encryption* based protocols, and *secure scalar product* based protocols. In what follows, we briefly review the fundamental cryptographic properties and the SMC protocols for PPRL for each category.

Note that SMC may be applied to secure information in other steps of the PPRL workflow. One example is that Al-Lawati et al. [2] adopt a secure set intersection protocol in Frugal Third Party Blocking, in order to find the set of common keywords shared by two data sources. Due to the expensive computation inflicted by encryption, SMC often is preceded by a privacy-preserving blocking step in the PPRL workflow, such as in [33, 53], which reduces the number of candidate record pairs to match through secure computation.

10.4.1 Commutative Encryption Based Protocols

An encryption scheme is commutative when a message is encrypted twice with two keys and the resulting ciphertext is independent of the order of encryptions. More formally, an encryption scheme $E(\cdot)$ is characterized as *commutative* if and only if, for any two keys e_1 and e_2 and for any message m , it holds that: (1) $E_{e_1}(E_{e_2}(m)) = E_{e_2}(E_{e_1}(m))$; (2) Encryption key e_i and its decryption key d_i are computable in polynomial time; and (3) $E_{e_i}(m)$ has the same value range. The commutative property applies to the decryption phase too. Examples of commutative encryption schemes include Pohlig-Hellman and SRA.

Du and Atallah [10] proposed a secure three-party protocol based on commutative encryption for approximate matching with a generic distance function f in the PIM (Private Information Matching) model. In short, in the PIM model, Alice has a string $x = x_1 \dots x_n$ and Bob has a database of strings $T = \{t_1, \dots, t_N\}$ and the length of each string t_i is n ; Alice wants to know the result of $\text{Match}(x, T)$. The requirement is that Bob should not know x or the result, and Alice should not learn more information than the reply from Bob. Given a generic function f , the

distance between two strings a and b can be measured by $\sum_{k=1}^n f(a_k, b_k)$. A two-party protocol was proposed to compute $f(x_k, t_{i,k})$ based on commutative encryption. The complete protocol employs a third party, Ursula, and the function protocol for a constructed $f_{i,k}(x_k, t_{i,k}) = f(x_k, t_{i,k}) + R_{i,k}$, in order to compute $f(x, t_i)$ for any i . As a result, Alice does not know the actual value for $f(x_k, t_{i,k})$ and Ursula does not know the actual value for the closest match. This protocol can be applied to evaluate $|a_k, b_k|$, $(a_k - b_k)^2$, and $\delta(a_k, b_k)$ and the communication cost is $O(m \times n \times N)$.

Li et al. [35] studied the problem of privacy-preserving group linkage (PPGL), where groups of individual records are linked if they are associated with the same entity. To prevent membership inference attacks, Alice and Bob negotiate a threshold θ and follow a protocol to match two groups of records, R and S . In the end, they only learn more than $|R|, |S|$, and a boolean result B (depending on whether $SIM(R, S) \geq \theta$). The authors proposed a two-party protocol for computing $SIM(R, S)$ with exact matching employed at the record level. Their solution is based on the secure set intersection protocol, proposed by Agrawal et al. [1], which is established on top of commutative encryption. The basic idea of the approach is to find the smallest integer k such that $SIM(R, S) \geq \theta$ and then follow [1] to find the intersection of all subsets of Alice and Bob that are of size k . If one k -combination is found in the intersection, the two groups are matched. Since Alice generates C_k^m subsets and Bob generates C_k^n subsets, this approach is only preferable when k is (1) very small, or (2) very close to m and n .

O’Keefe et al. [38] proposed several protocols for privacy preserving data linkage and extraction. The linkage protocol enables the exact matching of records in a secure manner across data sources. The extraction protocol extracts a cohort of individual data from a data source, without revealing the membership in the cohort. The design of the linkage protocol resembles the set intersection protocol in [1], with an addition of a third party to compute the actual intersection. Due to the commutative encryption schemes, the third party can match encrypted records correctly with negligible probability of hash collisions. As a result, each data source learns significantly less than in [1], in particular not enough to construct the intersection.

10.4.2 Homomorphic Encryption Based Protocols

Homomorphic encryption methods enable certain algebraic operations to be performed with ciphertext, which in the PPRL can be used to privately compute similarity functions between record pairs. Formally, given a homomorphic encryption scheme $E()$, ciphertexts $E(x)$ and $E(y)$, we are able to compute the *encrypted* $E(x \star y)$ without decryption, i.e., without knowing the plaintexts or private keys. Here \star represents an arithmetic operation such as addition and multiplication.

Additive homomorphic encryption, in particular the Paillier cryptosystem [39], has been widely adopted for privacy-preserving record linkage. Let $E_{pk}(\cdot)$ and $D_{pr}(\cdot)$ denote the encryption function with public key pk and the decryption function with

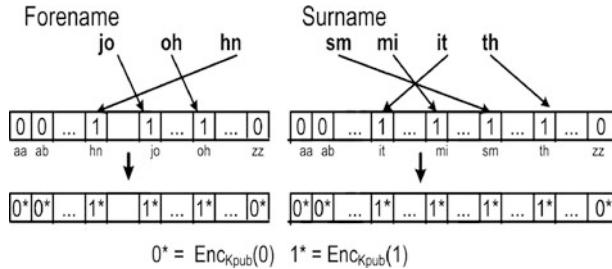


Fig. 10.5 Example of bitwise encryption by Kuzu et al. [33]

private key pr , respectively. The Paillier encryption has the following properties: (1) Given a constant k and a message m , the ciphertext of km can be computed by $E_{pk}(km) := k \times_h E_{pk}(m)$; (2) Given the ciphertexts $E_{pk}(m_1)$ and $E_{pk}(m_2)$, the encrypted sum can be computed by $E_{pk}(m_1 + m_2) := E_{pk}(m_1) +_h E_{pk}(m_2)$; (3) Given a message m , ciphertexts $c_1 = E_{pk}(m)$ and $c_2 = E_{pk}(m)$, $D_{pr}(c_1) = D_{pr}(c_2)$ but $c_1 \neq c_2$ with high probability.

The work of Kuzu et al. [33] employs the Paillier system to enable data sources, Alice and Bob, to compute secure similarity among encrypted records from the same blocks. The encrypted record similarity results are then sent to a third party, Charlie, to decrypt and execute the decision rule. In short, Charlie generates a Paillier public/private key pair; Alice and Bob then encrypt their records using the public key. Textual attributes, e.g., strings, can be represented as bit vectors, where each dimension represents a distinct 2-g, as illustrated in Fig. 10.5. Note that due to the probabilistic property of Paillier encryption, the ciphertexts of the same message, such as 0^* 's, are not same with high probability.

With the encrypted vector cardinality shared between data sources, Jaccard similarity between strings can be evaluated with encrypted bit vectors by either Alice or Bob, and the encrypted similarity can be sent to Charlie for decryption. On the other hand, numeric attributes and squares can be directly encrypted by data sources and Euclidean distance can be evaluated on encrypted data by any data source. Similarly, the actual Euclidean distance can be decrypted by Charlie and the id's of matched record pairs will be sent back to Alice and Bob. Furthermore, the authors proposed adaptations of the above protocols for *similarity blinding*, where the matching thresholds and actual similarities are perturbed with random values before sharing with Charlie. As a result, the third party cannot infer the exact value of record pair similarities when executing the decision rule.

Atallah et al. [3] proposed a two party protocol for computing *edit distance* between two sequences, such that neither party reveals anything about their private sequence to the other. The *weighted edit distance* between two sequences is the minimum cost of insertions, deletions, and substitutions required to transform one sequence to the other, where the cost of any operation is symbol-dependent. The secure protocol modifies the dynamic programming procedure for edit distance and the distance matrix is split across two parties, i.e., $M = M_A + M_B$. At each step, the

costs of the three operations are evaluated without disclosing at which position the minimum occurs. A *Blind-and-Permute* protocol based on homomorphic encryption is designed such that neither party knows the overall permutation and random vector. This approach allows for arbitrary values for the costs of insertion, deletion, and substitution. The longest common subsequence problem is a special case, where insertions and deletions have unit cost and substitutions are not considered. This approach has a considerable communication cost: one communication step is needed for each element in matrix M . Therefore, the approach is not well suited for performing linkage tasks on large databases.

In the work of Kantarcioğlu et al. [26], two data sources store encrypted data on an untrusted data site (DS) and a separate key holder site (KHS) manages the keys for encrypting data records/queries, as well as keys to decrypt the query results. A secure protocol, based on Paillier system, is proposed to execute equi-join queries by the DS: two records are linked if all joining attributes are exact matches. In order to prevent the KHS to learn the actual matching results, the authors design a matching scheme with a random vector and modular operation, and show that the false positive rate through the proposed scheme can be bounded by $1/(n-1)$ (where n is a large prime and the modular base). As a result, the KHS observes 0 for a match and a random value for a non-match. In order to avoid evaluating all possible record pairs from data sources, a hashing scheme based on k -anonymity is proposed and the DS can perform secure equi-join procedure within each bucket.

Li et al. [35] utilized homomorphic encryption schemes in linking groups of records above a threshold for exact matching and approximate matching at record level. For exact matching, the authors proposed a protocol similar to the set intersection protocol in [18], with perturbation with random numbers in order to hide the actual value of the intersection of two groups. The approximate matching protocol employs another threshold ρ at the record level, and those record pairs with similarities above ρ are considered as matches. To avoid attacks by faking group records, this protocol securely computes the number of records from Alice which are linked to at least one record from Bob, and vice versa. Then, the smaller of the two will be compared with the minimum intersection size to make the linking decision.

10.4.3 Secure Scalar Product Protocols

Secure scalar product protocols belong to a generic framework that enables one party to find out the scalar product of several parties' private vectors. For example, the Euclidean distance $d'(\cdot, \cdot)$ between the vectors \bar{x} and \bar{y} can be approximated as follows:

$$d'(\bar{x}, \bar{y})^2 = \|\bar{x} - \bar{y}\|_2^2 = \|\bar{x}\|_2^2 + \|\bar{y}\|_2^2 - 2 \langle \bar{x}, \bar{y} \rangle \quad (10.1)$$

where only the scalar product $\langle \bar{x}, \bar{y} \rangle$ is computed in a secure way by the SMC protocol. This protocol is typically used by several PPRL techniques to privately evaluate the similarity between data records that have been previously transformed into vectors in the data transformation step.

Ravikumar et al. [41] proposed a stochastic scalar product protocol for secure computation of several standard distance metrics, such as TFIDF, SoftTFIDF, and the Euclidean distance. To convert a string s to a real-valued vector, s is first broken into a set of tokens, i.e., words. Each token w in s is given a numeric weight, $weight(w, s) = \log(TF_{w,s} + 1) \times \log(IDF_w)$, where $TF_{w,s}$ represents the term-frequency of w in s and IDF_w represents the inverse of the fraction of strings in the corpus that contain w . Similarity metrics between two strings r and s can then be evaluated based on the weighted feature vectors V_r and V_s . The scalar product protocol requires L1 normalization of each party's feature vector. And each party samples its feature vector with probability $\frac{V(i)}{Z}$ for every word w_i , where Z is the normalizing factor. A secure set intersection protocol is needed to find the intersection of sample sets T_r and T_s . An unbiased estimate of the scalar product can then be derived: $\frac{|T_r \cap T_s|}{numSamples} \times Z_r \times Z_s$. The addition information disclosed beyond the intersection protocol is the normalization factors Z_r and Z_s .

Yakout et al. [53] proposed a two-party private record linkage protocol with approximate matching. The authors assume that data records are first transformed to numeric vectors by each party, which can be done using techniques described in [43]. In the first protocol, the numeric vectors are projected in the complex plane and likely linked pairs can be computed between complex numbers. This step reduces the number of candidate record pairs to be evaluated in the second protocol, where the scalar product is derived securely without using cryptographic operations. In short, party A generates linearly independent random vectors and random scalars to hide A 's data record from party B . Similarly, B receives perturbed data from A and perturbs/contributes its own record with orthogonal random vectors and random scalars. The output of the protocol is that B obtains the scalar product without learning A 's record. As pointed out by the authors, B also learns the k -dimensional hyperplane that contains A 's record, and they suggest that a larger k would possibly increase A 's privacy. Moreover, A can choose a relatively innocuous hyperplane when generating random vectors.

Du and Atallah [10] proposed a Private Information Matching (PIM) model and designed a three-party secure scalar product for computing the square distance between two messages. For $x = x_1 \dots x_n$ and each $t_i = t_{i,1} \dots t_{i,n}$ in T , the square distance can be calculated by: $\sum_{k=1}^n (x_k - t_{i,k})^2 = (-2x_1, \dots, -2x_n, 1) \cdot (t_{i,1}, \dots, t_{i,n}, \sum_{k=1}^n t_{i,k}^2) + \sum_{k=1}^n x_k^2$. Since x is fixed in the PIM model, $\sum_{k=1}^n x_k^2$ is constant. The closest match can be found by computing the scalar product between $(-2x_1, \dots, -2x_n, 1)$ and $(t_{i,1}, \dots, t_{i,n}, \sum_{k=1}^n t_{i,k}^2)$. The secure scalar product protocol involves random vectors in order to disguise the query and private records, and random numbers in order to disguise the matching results from the third party. The communication cost is $O(n)$ for each record pair.

10.5 Hybrid Approaches

Hybrid approaches combine blocking/indexing techniques with SMC protocols with the aim of reducing the SMC cost for evaluating record pairs. In this section, we briefly review those approaches and the blocking techniques they adopt.

For generic record linkage applications, the blocking/indexing step plays a central role from both the scalability and the utility perspective. Suppose two databases D_A and D_B , containing N_A and N_B records respectively, are directly fed into the matching step. It leads to $N_A \times N_B$ pairwise record comparisons, which is likely to become a major performance bottleneck for the entire process. This expensive matching operation between all record pairs is not necessary in general. In fact, due to apparent differences in attribute values, such as gender and age, only a small fraction of the complete set of record pairs are likely to be matches, while the majority correspond to non-matches. Motivated by this observation, researchers have been developing techniques to reduce the unnecessary computations in the past 20 years. Typically, records are grouped into blocks and only those who fall into the same block will be compared for matching.

Although the use of *blocking* enables the linkage of large databases, it also introduces a trade-off between the computational cost and the quality of matching. In principle, a large number of small blocks is beneficial for reducing the number of pairwise comparisons. However, it also increases the chance of missing true matching records, due to the fact that similar records could be placed in different blocks. This utility loss is particularly evident in real-world databases, where data records often contain erroneous information, such as typos and outdated data, and matching records may be misplaced into different blocks as a result. On the other hand, the use of large blocks reduces the chance of losing true matching records; however, the computational cost for comparing the records increases. For a more detailed description and comparison between blocking techniques, we point the interested readers to the surveys presented in [4, 8, 37].

In exiting PPRL solutions, the blocking step is usually performed in two ways: (1) on *original data records*, or (2) on *transformed data records*. We identify four major categories for the blocking techniques adopted in the PPRL literature, i.e., *standard blocking*, *sorted neighborhood*, *mapping* and *clustering*. In the standard blocking, records are associated to a key value and only those records having the same key are placed in the same block and further considered as candidate matches. The sorted neighborhood approach sorts all records according to a chosen key, and only those record pairs that appear close in the sorted list are preserved for matching. For those approaches where the blocking is performed via mapping, a combination of record attributes is used to define a blocking key. The similarity between the key values reflects the similarity between the original records, which enables the construction of blocks that are more robust with respect to the noise in the data and more effective in retrieving the true matching records. The intuition is that non-matching records will produce different key values and they will be placed in different blocks, so that the direct comparison of such records will be avoided and

they will be marked as non-matching records. Last, in the clustering based blocking, records are clustered according to specific similarity metrics and records belonging to the same cluster are placed in the same block.

10.5.1 Standard Blocking

Standard blocking techniques group records in the same block when they have an identical *blocking key*. If b denotes the number of blocks (all of the same size), then the number of pair comparisons is reduced to $O(N_A \times N_B/b)$. However, this result is hard to achieve in real applications, since often the blocks have different size and therefore the running time is dominated by the largest block. Below, we discuss some recent PPRL approaches that implement this blocking technique.

The work proposed by Al-Lawati et al. [2] considers three different blocking techniques, where the blocking key for each record is obtained via a hash function. This approach requires the presence of a third party in charge of matching the records within the same block. The records are encoded using a secure hashing transformation and the hash signatures determine the block in which each record will be placed. In the first blocking schema proposed in [2] (*simple blocking*), for each record's attribute a hash function is applied on the original record and the result is placed in the block indexed by the hash signature. As the technique is designed, it is possible that the similarity of a pair of records could be computed several times if records appear together in multiple blocks, leading to poor performance in the case of a high ratio of common blocks per record. The second blocking approach, called *record-aware blocking*, solves this problem by coupling an id to the hash signature of each record. Both these approaches heavily involve the third party, resulting in a high communication cost due to the transmission of all the blocks. To reduce this issue, the authors proposed a frugal third party blocking strategy, which uses a SMC protocol to limit the transfer of only the common blocks between the two parties.

Kantarcioglu et al. [26] proposed a PPRL technique where the blocking step achieves privacy by guaranteeing k -anonymity for the records in the blocks. The original records are first de-identified by using k -anonymity, and such values are used as hashed keys to partition the encrypted data into blocks containing at least k records each. Then, the records are matched using a secure equi-join protocol.

Inan et al. [24] proposed a hybrid model for record linkage based on a two party SMC protocol. The blocking of the records is achieved with the help of multi-dimensional tree index data structures (i.e., BSP-tree, adaptive kd-tree, and R*-tree). The construction of the data partitions forming the blocks is performed by enforcing differential privacy. The idea consists in perturbing the count of the record in each partition, resulting in removing some real records (negative noise) and introducing fake records (positive noise). In this way, the authors protect the presence of any individual record in the original databases. Finally, the comparison of the records is performed via a SMC protocol among the data holders, by considering the pairs of overlapping blocks, as shown in Fig. 10.6.

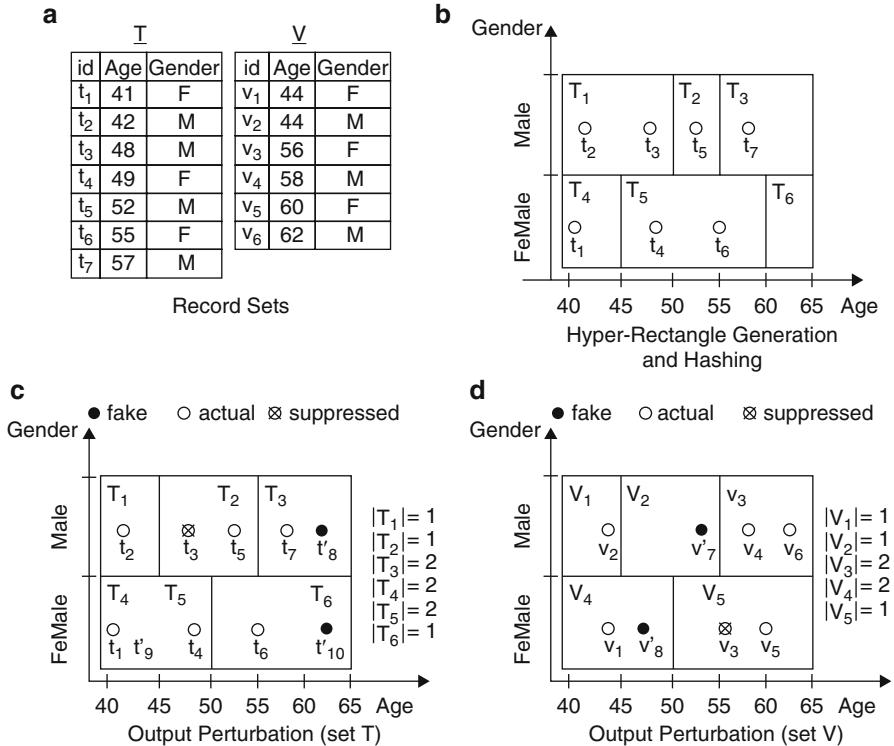


Fig. 10.6 Performing blocking on datasets: (a) Original datasets T and V ; (b) Block decomposition using hyper-rectangles for T ; (c) Block perturbation for T ; (d) Block perturbation for V . The candidate matching pairs tested in the SMC part are limited to the pair of overlapping blocks: $(T_1, V_1), (T_1, V_2), \dots, (T_5, V_5)$ (example from Inan et al. [24])

Recently, Kuzu et al. [33] proposed a blocking procedure which extends the work in [24]. In this case, the records are partitioned according to some public identifiers provided by a third party. To ensure the privacy of each individual record in the blocks, the differential privacy mechanism is employed. The authors showed how to calibrate the amount of perturbation introduced by the differential privacy mechanism, by shifting the mean of the noise distribution to minimize the number of expected comparisons and optimize the number of suppressed records. The comparison between records within the same block is achieved via a SMC protocol.

10.5.2 Sorted Neighborhood Approach

The sorted neighborhood approach consists of the following three steps. Firstly, a key is created for each record by extracting relevant attributes or portions of attributes. All records in the database are subsequently sorted according to their key

values. Lastly, a fixed-size window is moved through the list of sorted records, and only record pairs that co-appear inside the window are compared in the matching step. The assumption of this approach is that matching records should be close in the sorted list. As a result, the effectiveness of the approach is highly dependent on the quality of the key chosen to sort the records.

Yakout et al. [53] proposed a private protocol for computing likely-linked pairs, followed by securely matching the candidate pairs. To start, each party converts her record vectors to complex numbers by taking the first elements of discrete Fourier transform (DFT). Secondly, record pairs close in the complex plane are selected by moving a vertical slab with fixed width along the real axis. The authors show that this approach leads to no false negatives, as vectors close to each other will be also close in the complex plane. As for privacy, the authors claim that it is impractical to infer the original vectors from the corresponding complex numbers, as they reveal little information, i.e., only the first element of DFT, about the original records.

10.5.3 Mapping

Mapping approaches use encoding techniques and hashing functions in order to determine which records should be blocked together. These techniques tend to be more robust to typos and other errors in the original records, when compared with standard blocking techniques.

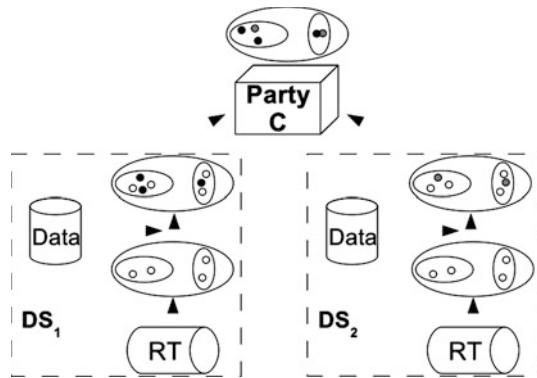
Karakasidis and Verykios [27] proposed a two party protocol for records linkage, where the secure blocking step is based on mapping. The original records are transformed using phonetic encoding, using an encoding map for each attribute, and adding fake records to increase security. The resulting records are hashed using a shared hashing function. The matching of the records is performed via a SMC protocol between the two parties.

Durham [11] proposed a PPRL framework, where the blocking step is performed using locality-sensitive hashing functions (LSH) [20]. The transformed records, encoded using composite Bloom filters, are blocked using several hash functions, chosen uniformly, independently at random from a family of hash functions, which are locally sensitive to a distance metric. Durham proposed to use Hamming-based Locality Sensitive Hashing functions in combination with composite Bloom filters. The hash functions are used to project the Bloom filter representation of the records on a fixed number of bits. Then, the records having the same bits pattern are placed in the same block. The author showed that the mapped records can be matched in an approximate way, and the encoding prevents cryptanalysis attacks.

10.5.4 Clustering

Clustering approaches are often employed when the original records are embedded into a vector space, where the similarity distance between vectors reflects the similarity between original records. The blocking step takes advantage of the

Fig. 10.7 Private blocking via clustering (example from [28])



geometrical property of the embedded space to group the mapped records and discard the pairs of records that do not represent a match.

Karakasidis and Verykios [28] proposed a three-party PPRL approach where the records are blocked according to a clustering criterion. The data owners share a reference table that is used to initialize the clusters. The data owners create a set consisting of unique elements from the shared reference table, and form initial clusters containing at least k elements of the reference set each. After this initial step, data owners form their own clusters using the nearest neighbor clustering criterion, where the similarity measure employed is the Dice coefficient. The similarity comparison is performed between the records in the original data and those in the reference table (RT). Each data owner assigns a set blocking key for each respective cluster. These clusters are then sent to the trusted third party, which computes the candidate matching pairs, as shown in Fig. 10.7. This approach supports approximate matching and provides privacy guarantees in the form of k -anonymity.

Pang et al. [40] developed a PPRL approach, where the data owners use a set of reference strings for matching records representing patients' names. The data owners start by computing the distance of their strings with respect to the reference set, and send their information to the third party. For each pair of records, the third party uses the triangulation property of the distance measure employed, in order to estimate if the original string could represent a match. The matching decision can be performed with different distance measures, for example edit distance, and using a threshold value the third party is able to identify records that match in an approximate way. To reduce the number of pairs in the final comparison, the nearest neighbor clustering approach is applied as a blocking step. In this approach, the distance information with respect to the reference set of strings is revealed to the third party, while the original private data is never shared among the parties.

10.6 Challenges and Future Research Directions

Despite the continuous development in designing privacy-preserving record linkage approaches to provide effective and efficient solutions with privacy guarantees, several challenges remain unsolved, creating significant opportunities for investigating future research directions. In this section, we discuss the main future research questions for privacy-preserving record linkage in the biomedical domain.

Privacy Aspect Most of the PPRL techniques that we considered in the chapter provide privacy guarantee using encryption or k -anonymity. However, such notions could lead to privacy leakage when an adversary has external/background knowledge regarding patients' data. Furthermore, the large amount of personal data publicly available nowadays (i.e., Facebook profiles, public register data, etc.) is elevating the risk of potential privacy breaches. To address this problem, the recently proposed notion of differential privacy has been employed by a few PPRL approaches. Despite the strong privacy guarantee provided by this technique, differential privacy is limited to privately answering only statistical queries. Strong privacy techniques tailored to the record linkage process should be developed in the future.

Medical records present a high level of complexity: text report, snip of DNA sequences, name and age, are only a few examples of information related to a patient's record. Due to the nature of such attributes, some of them are more sensitive than the other. For example, many patients could have the same name but the DNA sequence could uniquely identify the patient. Therefore, the use of a universal privacy notion among all the attributes could result either in poor utility, due to a stringent privacy guarantee, or in leakage of sensitive information, in the case of a weaker privacy notion. Since the attributes are not equally sensitive, more research is needed to design PPRL protocols that allow attribute-specific privacy notions.

Recently, we can observe a high degree of decentralization for medical data. Patient's medical records are often distributed among a multitude of organizations. This situation poses additional privacy challenges in linking the records from a large number of data sources. In this setting, multiple parties could collude, increasing the ability of the adversary to breach the privacy of individual users. Most existing PPRL protocols typically deal with linkage between two data owners. Therefore, more research is necessary for designing record linkage solutions to achieve privacy in a highly-distributed setting.

Linkage Quality Aspect Despite the fact that many PPRL solutions allow for approximate matching between records, they are mostly limited only to the string data type. In fact, new technologies in the medical domain (e.g., medical image processing, DNA sequencing, etc.) enable the collection of more and complex information about each patient, which could be potentially considered for record linkage purposes. Future research is needed to develop privacy-preserving protocols for computing approximate comparison metrics for those complex data types, incorporating the rich information available in a patient's profile nowadays.

An emerging research direction in privacy-preserving record linkage is to allow human interaction with the automated process, in order to improve the final matching utility [31]. This can be achieved by human experts who can fine tune the matching results and manage the uncertainty and its propagation into subsequent analyses. However, we identify several new challenges introduced by this interactive approach, such as *data disclosure risks* from the matching results and *inference risks* from intermediate tuning results. Future research may investigate methods to securely perform the human interaction step in record linkage, as well as the incorporation of it into the overall PPRL framework.

Scalability Aspect Most of the existing PPRL techniques employ an indexing/blocking step to improve the efficiency of record matching. However, only few works [24, 33, 36] propose collaborative blocking schemes between parties, improving the utility of blocking by taking into account the dataset information without compromising privacy. This represents an interesting and important research direction for designing more scalable PPRL solutions.

10.7 Conclusion

In this chapter, we examined the problem of privacy preserving record linkage (PPRL). We reviewed the most recent research works on PPRL, many of which have particular focus on medical data applications. A taxonomy of existing works, regarding *privacy*, *scalability*, and *linkage quality*, was provided to facilitate the understanding and implementation of these protocols. We summarized a conceptual framework for PPRL and illustrated the privacy assurance that the techniques in literature give at each step of the PPRL process. Toward the end, we identified the challenges of PPRL applications in the biomedical domain and we pointed out some possible promising future research directions.

References

1. Agrawal, R., Evfimievski, A., Srikant, R.: Information sharing across private databases. In: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, SIGMOD '03, pp. 86–97. ACM, New York, NY (2003). doi:[10.1145/872757.872771](https://doi.org/10.1145/872757.872771). <http://www.doi.acm.org/10.1145/872757.872771>
2. Al-Lawati, A., Lee, D., McDaniel, P.: Blocking-aware private record linkage. In: Proceedings of the 2nd International Workshop on Information Quality in Information Systems, IQIS '05, pp. 59–68. ACM, New York, NY (2005). doi:[10.1145/1077501.1077513](https://doi.org/10.1145/1077501.1077513). <http://www.doi.acm.org/10.1145/1077501.1077513>
3. Atallah, M.J., Kerschbaum, F., Du, W.: Secure and private sequence comparisons. In: Proceedings of the 2003 ACM Workshop on Privacy in the Electronic Society, WPES '03, pp. 39–44. ACM, New York, NY (2003). doi:[10.1145/1005140.1005147](https://doi.org/10.1145/1005140.1005147). <http://www.doi.acm.org/10.1145/1005140.1005147>

4. Baxter, R., Christen, P., Churches, T.: A comparison of fast blocking methods for record linkage. In: ACM SIGKDD, vol. 3, pp. 25–27. Citeseer (2003)
5. Bloom, B.H.: Space/time trade-offs in hash coding with allowable errors. Commun. ACM **13**(7), 422–426 (1970). doi:[10.1145/362686.362692](https://doi.org/10.1145/362686.362692). <http://www.doi.acm.org/10.1145/362686.362692>
6. Bonomi, L., Xiong, L., Chen, R., Fung, B.C.: Frequent grams based embedding for privacy preserving record linkage. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM ’12, pp. 1597–1601. ACM, New York, NY (2012). doi:[10.1145/2396761.2398480](https://doi.org/10.1145/2396761.2398480). <http://www.doi.acm.org/10.1145/2396761.2398480>
7. Christen, P.: A comparison of personal name matching: Techniques and practical issues. In: Sixth IEEE International Conference on Data Mining Workshops, 2006. ICDM Workshops 2006, pp. 290–294 (2006). doi:[10.1109/ICDMW.2006.2](https://doi.org/10.1109/ICDMW.2006.2)
8. Christen, P.: A survey of indexing techniques for scalable record linkage and deduplication. IEEE Trans. Knowl. Data Eng. **24**(9), 1537–1555 (2012). doi:[10.1109/TKDE.2011.127](https://doi.org/10.1109/TKDE.2011.127). <http://www.dx.doi.org/10.1109/TKDE.2011.127>
9. Churches, T., Christen, P.: Some methods for blindfolded record linkage. BMC Med. Inform. Decis. Mak. **4**(1), 9 (2004). doi:[10.1186/1472-6947-4-9](https://doi.org/10.1186/1472-6947-4-9). <http://www.biomedcentral.com/1472-6947/4/9>
10. Du, W., Atallah, M.J.: Protocols for secure remote database access with approximate matching. In: Ghosh, A. (ed.) E-Commerce Security and Privacy. Advances in Information Security, vol. 2, pp. 87–111. Springer, Berlin (2001). doi:[10.1007/978-1-4615-1467-1_6](https://doi.org/10.1007/978-1-4615-1467-1_6). http://www.dx.doi.org/10.1007/978-1-4615-1467-1_6
11. Durham, E.A.: A framework for accurate, efficient private record linkage. Ph.D. Thesis (2012)
12. Durham, E., Xue, Y., Kantarcioglu, M., Malin, B.: Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. Inf. Fusion **13**(4), 245–259 (2012). doi:[10.1016/j.inffus.2011.04.004](https://doi.org/10.1016/j.inffus.2011.04.004). <http://www.dx.doi.org/10.1016/j.inffus.2011.04.004>
13. Durham, E.A., Kantarcioglu, M., Xue, Y., Toth, C., Kuzu, M., Malin, B.: Composite bloom filters for secure record linkage. IEEE Trans. Knowl. Data Eng. **99**(preprints), 1 (2013). <http://www.doi.ieeecomputersociety.org/10.1109/TKDE.2013.91>
14. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) Automata, Languages and Programming. Lecture Notes in Computer Science, vol. 4052, pp. 1–12. Springer, Berlin, Heidelberg (2006). doi:[10.1007/11787006_1](https://doi.org/10.1007/11787006_1). http://www.dx.doi.org/10.1007/11787006_1
15. Dwork, C.: Differential privacy: a survey of results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds.) Theory and Applications of Models of Computation. Lecture Notes in Computer Science, vol. 4978, pp. 1–19. Springer, Berlin, Heidelberg (2008). doi:[10.1007/978-3-540-79228-4_1](https://doi.org/10.1007/978-3-540-79228-4_1). http://www.dx.doi.org/10.1007/978-3-540-79228-4_1
16. Elmagarmid, A., Ipeirotis, P., Verykios, V.: Duplicate record detection: a survey. IEEE Trans. Knowl. Data Eng. **19**(1), 1–16 (2007). doi:[10.1109/TKDE.2007.250581](https://doi.org/10.1109/TKDE.2007.250581)
17. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. J. Am. Stat. Assoc. **64**(328), 1183–1210 (1969)
18. Freedman, M., Nissim, K., Pinkas, B.: Efficient private matching and set intersection. In: Cachin, C., Camenisch, J. (eds.) Advances in Cryptology - EUROCRYPT 2004. Lecture Notes in Computer Science, vol. 3027, pp. 1–19. Springer, Berlin, Heidelberg (2004). doi:[10.1007/978-3-540-24676-3_1](https://doi.org/10.1007/978-3-540-24676-3_1). http://www.dx.doi.org/10.1007/978-3-540-24676-3_1
19. Getoor, L., Machanavajjhala, A.: Entity resolution: theory, practice & open challenges. Proc. VLDB Endow. **5**(12), 2018–2019 (2012). doi:[10.14778/2367502.2367564](https://doi.org/10.14778/2367502.2367564). <http://www.dx.doi.org/10.14778/2367502.2367564>
20. Gionis, A., Indyk, P., Motwani, R.: Similarity search in high dimensions via hashing. In: Proceedings of the 25th International Conference on Very Large Data Bases, VLDB ’99, pp. 518–529. Morgan Kaufmann Publishers Inc., San Francisco, CA (1999). <http://www.dl.acm.org/citation.cfm?id=645925.671516>

21. Goldreich, O.: Foundations of Cryptography: Volume 2, Basic Applications. Cambridge University Press, New York, NY (2004)
22. Hall, R., Fienberg, S.E.: Privacy-preserving record linkage. In: Proceedings of the 2010 International Conference on Privacy in Statistical Databases, PSD'10, pp. 269–283. Springer, Berlin, Heidelberg (2010). <http://www.dl.acm.org/citation.cfm?id=188848.188878>
23. Hjaltason, G., Samet, H.: Properties of embedding methods for similarity searching in metric spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(5), 530–549 (2003)
24. Inan, A., Kantarcioglu, M., Ghinita, G., Bertino, E.: Private record matching using differential privacy. In: Proceedings of the 13th International Conference on Extending Database Technology, EDBT '10, pp. 123–134. ACM, New York, NY (2010). doi:[10.1145/1739041.1739059](https://doi.org/10.1145/1739041.1739059) <http://www.doi.acm.org/10.1145/1739041.1739059>
25. Jiang, W., Clifton, C.: A secure distributed framework for achieving k-anonymity. *VLDB J.* **15**(4), 316–333 (2006). doi:[10.1007/s00778-006-0008-z](https://doi.org/10.1007/s00778-006-0008-z). <http://www.dx.doi.org/10.1007/s00778-006-0008-z>
26. Kantarcioglu, M., Jiang, W., Malin, B.: A privacy-preserving framework for integrating person-specific databases. In: Domingo-Ferrer, J., Saygn, Y. (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science, vol. 5262, pp. 298–314. Springer, Berlin, Heidelberg (2008). doi:[10.1007/978-3-540-87471-3_25](https://doi.org/10.1007/978-3-540-87471-3_25). http://www.dx.doi.org/10.1007/978-3-540-87471-3_25
27. Karakasidis, A., Verykios, V.S.: Secure blocking + secure matching = secure record linkage. *J. Comput. Sci. Eng.* **5**(3), 223–235 (2011)
28. Karakasidis, A., Verykios, V.S.: Reference table based k-anonymous private blocking. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing, SAC '12, pp. 859–864. ACM, New York, NY (2012). doi:[10.1145/2245276.2245444](https://doi.org/10.1145/2245276.2245444). <http://www.doi.acm.org/10.1145/2245276.2245444>
29. Karakasidis, A., Verykios, V.S., Christen, P.: Fake injection strategies for private phonetic matching. In: Proceedings of the 6th International Conference, and 4th International Conference on Data Privacy Management and Autonomous Spontaneous Security, DPM'11, pp. 9–24. Springer, Berlin, Heidelberg (2012). doi:[10.1007/978-3-642-28879-1_2](https://doi.org/10.1007/978-3-642-28879-1_2). http://www.dx.doi.org/10.1007/978-3-642-28879-1_2
30. Kirsch, A., Mitzenmacher, M.: Less hashing, same performance: building a better bloom filter. *Random Struct. Algoritm.* **33**(2), 187–218 (2008). doi:[10.1002/rsa.v33:2](https://doi.org/10.1002/rsa.v33:2). <http://www.dx.doi.org/10.1002/rsa.v33:2>
31. Kum, H.C., Krishnamurthy, A., Machanavajjhala, A., Reiter, M.K., Ahalt, S.: Privacy preserving interactive record linkage (ppirl). *J. Am. Med. Inform. Assoc.* **21**(2), 212–220 (2014). doi:[10.1136/amiajnl-2013-002165](https://doi.org/10.1136/amiajnl-2013-002165)
32. Kuzu, M., Kantarcioglu, M., Durham, E., Malin, B.: A constraint satisfaction cryptanalysis of bloom filters in private record linkage. In: Fischer-Hübner, S., Hopper, N. (eds.) Privacy Enhancing Technologies. Lecture Notes in Computer Science, vol. 6794, pp. 226–245. Springer, Berlin, Heidelberg (2011). doi:[10.1007/978-3-642-22263-4_13](https://doi.org/10.1007/978-3-642-22263-4_13). http://www.dx.doi.org/10.1007/978-3-642-22263-4_13
33. Kuzu, M., Kantarcioglu, M., Inan, A., Bertino, E., Durham, E., Malin, B.: Efficient privacy-aware record integration. In: Proceedings of the 16th International Conference on Extending Database Technology, EDBT '13, pp. 167–178. ACM, New York, NY (2013). doi:[10.1145/2452376.2452398](https://doi.org/10.1145/2452376.2452398). <http://www.doi.acm.org/10.1145/2452376.2452398>
34. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. *Tech. Rep.* 8 (1966)
35. Li, F., Chen, Y., Luo, B., Lee, D., Liu, P.: Privacy preserving group linkage. In: Proceedings of the 23rd International Conference on Scientific and Statistical Database Management, SSDBM'11, pp. 432–450. Springer, Berlin, Heidelberg (2011). <http://www.dl.acm.org/citation.cfm?id=2032397.2032432>
36. Mohammed, N., Fung, B., Debbabi, M.: Anonymity meets game theory: secure data integration with malicious participants. *VLDB J.* **20**(4), 567–588 (2011). doi:[10.1007/s00778-010-0214-6](https://doi.org/10.1007/s00778-010-0214-6). <http://www.dx.doi.org/10.1007/s00778-010-0214-6>

37. Nin, J., Muntes-Mulero, V., Martinez-Bazan, N., Larriba-Pey, J.L.: On the use of semantic blocking techniques for data cleansing and integration. In: Proceedings of the 11th International Database Engineering and Applications Symposium, IDEAS '07, pp. 190–198. IEEE Computer Society, Washington, DC (2007). doi:[10.1109/IDEAS.2007.36](https://doi.org/10.1109/IDEAS.2007.36). <http://www.dx.doi.org/10.1109/IDEAS.2007.36>
38. O'Keefe, C.M., Yung, M., Gu, L., Baxter, R.: Privacy-preserving data linkage protocols. In: Proceedings of the 2004 ACM Workshop on Privacy in the Electronic Society, WPES '04, pp. 94–102. ACM, New York, NY (2004). doi:[10.1145/1029179.1029203](https://doi.org/10.1145/1029179.1029203). <http://www.doi.acm.org/10.1145/1029179.1029203>
39. Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Proceedings of the 17th International Conference on Theory and Application of Cryptographic Techniques, EUROCRYPT'99, pp. 223–238. Springer, Berlin, Heidelberg (1999). <http://www.dl.acm.org/citation.cfm?id=1756123.1756146>
40. Pang, C., Gu, L., Hansen, D., Maeder, A.: Privacy-preserving fuzzy matching using a public reference table. In: Intelligent Patient Management, pp. 71–89. Springer, Berlin, Heidelberg (2009)
41. Ravikumar, P., Cohen, W.W., Fienberg, S.E.: A secure protocol for computing string distance metrics. In: IEEE ICDM Workshop on Privacy and Security Aspects of Data Mining (2004)
42. Salton, G. (ed.): Automatic Text Processing. Addison-Wesley Longman Publishing Co Inc., Boston, MA (1988)
43. Scannapieco, M., Figotin, I., Bertino, E., Elmagarmid, A.K.: Privacy preserving schema and data matching. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, SIGMOD '07, pp. 653–664. ACM, New York, NY (2007). doi:[10.1145/1247480.1247553](https://doi.acm.org/10.1145/1247480.1247553). <http://doi.acm.org/10.1145/1247480.1247553>
44. Schneier, B.: Applied Cryptography: Protocols, Algorithms, and Source Code in C, 2nd edn. Wiley, New York, NY (1995)
45. Schnell, R., Bachteler, T., Reiher, J.: Privacy-preserving record linkage using bloom filters. BMC Med. Inform. Decis. Mak. **9**(1), 41 (2009). doi:[10.1186/1472-6947-9-41](https://doi.org/10.1186/1472-6947-9-41). <http://www.biomedcentral.com/1472-6947/9/41>
46. Schnell, R., Bachteler, T., Reiher, J.: A novel error-tolerant anonymous linking code. In: Working Paper WP-GRLC-2011-02, German Record Linkage Center, Duisburg (2011)
47. Sweeney, L.: Weaving technology and policy together to maintain confidentiality. J. Law Med. Ethics **25**(2–3), 98–110 (1997). doi:[10.1111/j.1748-720x.1997.tb01885.x](https://doi.org/10.1111/j.1748-720x.1997.tb01885.x). <http://www.dx.doi.org/10.1111/j.1748-720x.1997.tb01885.x>
48. Sweeney, L.: K-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowledge Based Syst. **10**(5), 557–570 (2002). doi:[10.1142/S0218488502001648](https://doi.org/10.1142/S0218488502001648). <http://dx.doi.org/10.1142/S0218488502001648>
49. The European Parliament and the council of the European Union: EU Directive 95/46/EC. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML> (1995). Accessed 31 July 2014
50. Treptin, S.: Privacy-preserving string comparisons in record linkage systems: a review. Inf. Secur. J. Glob. Perspect. **17**(5–6), 253–266 (2008). doi:[10.1080/19393550802492503](https://doi.org/10.1080/19393550802492503). <http://www.dx.doi.org/10.1080/19393550802492503>
51. U.S. Department of Health and Human Services: HIPAA privacy rule. <http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/> (2002). Accessed 31 July 2014
52. Vatsalan, D., Christen, P., Verykios, V.S.: A taxonomy of privacy-preserving record linkage techniques. Inf. Syst. **38**(6), 946–969 (2013). doi:[10.1016/j.is.2012.11.005](https://doi.org/10.1016/j.is.2012.11.005). <http://www.dx.doi.org/10.1016/j.is.2012.11.005>
53. Yakout, M., Atallah, M.J., Elmagarmid, A.: Efficient and practical approach for private record linkage. J. Data Inf. Qual. **3**(3), 5:1–5:28 (2012). doi:[10.1145/2287714.2287715](https://doi.org/10.1145/2287714.2287715). <http://doi.acm.org/10.1145/2287714.2287715>

Chapter 11

Application of Privacy-Preserving Techniques in Operational Record Linkage Centres

James H. Boyd, Sean M. Randall, and Anna M. Ferrante

Abstract Record linkage is the process of bringing together data relating to the same individual within and between different datasets. These integrated datasets provide diverse and rich resources for researchers without the cost associated with additional data collection. By their nature, record linkage systems deal with large volumes of data and require complex organizational and technical infrastructure. Bringing together information from different sources often requires many different organizations to collaborate and share data, which presents challenges around data privacy and confidentiality. Various processes and protocols have been developed to protect the privacy of individuals during the record linkage process. These include data governance procedures covering people, processes and information technology, role separation and restricted data flows. Combinations of these are used to mitigate risks to privacy by limiting access to certain information. In addition, privacy-preserving record linkage techniques can be utilized to further reduce the risk to privacy, by removing all personal identifying information from linkage protocols. This chapter reviews current practices, processes and developments for maintaining security and privacy as applied in existing record linkage centres. Models for role separation and data flows are outlined and evaluated, and requirements for an effective privacy-preserving record linkage protocol are described.

11.1 Introduction

Record linkage involves identifying records which belong to the same individual, across datasets. This process is widely used to enable and enhance research in health and health services. Medical record linkage brings together administrative records from hospital and emergency collections, primary care facilities, birth, death and disease registries [20, 24]. Increasingly, record linkage methods are used to connect health records with data from other sectors including education, child protection and

J.H. Boyd (✉) • S.M. Randall • A.M. Ferrante

Centre for Data Linkage (CDL), Curtin University of Technology, Perth, WA, Australia

e-mail: j.boyd@curtin.edu.au; sean.randall@curtin.edu.au; a.ferrante@curtin.edu.au

criminal justice in order to investigate the social determinants of health, as well as other aspects of human function and well-being [8, 12, 13, 38].

Health and social service systems are complex with many interactions and relationships. Taking a holistic approach recognizes that many interacting factors can influence individual parts of the system. The record linkage process creates a chronological sequence of events for all individuals in a population. This provides researchers with cost-effective longitudinal data resources for an entire population.

In the absence of a unique person identifier, record linkage is carried out using personally identifying information such as names, dates of birth and addresses. As these identifiers can change and/or can include errors, probabilistic statistical methods are frequently employed to ensure high quality links [10].

11.1.1 Record Linkage Research Infrastructure

A significant investment in medical record linkage infrastructure has occurred in England [17], Scotland [29], Wales [16], Canada [37] and Australia [5], over the last 30 years. Dedicated record linkage centres with secure environments and specialized linkage personnel have been established in each of these countries [5]. Within each of the facilities, data is routinely linked to support health research and for on-going reporting, policy development and service design by government [32]. Research using linked data has led to changes in policy and health service delivery, which have resulted in improved services, reduced cost and improved outcomes [7, 21]. A summary of selected centres which undertake routine medical record linkage tasks is provided in Box 1 (below).

Box 1 International Record Linkage Centres

Australia

Centre for Data Linkage (CDL) [5]. Established in 2009, the CDL provides linkage services at a national level, linking datasets across different Australian states and territories, and linking these to external datasets or to other national data collections. Currently working on a project basis only, the CDL has undertaken a number of proof-of-concept linkages, each utilizing tens of millions of records.

Centre for Health Record Linkage (CHeReL) [31]. Established in 2006, the Centre for Health Record Linkage provides record linkage services for New South Wales and the Australian Capital Territory. The linkage system consists of almost 100 million records for over 11 million people and has provided data for over 150 projects.

(continued)

Data Linkage Branch (DLB), Western Australia [20]. Record linkage has been used for health and medical research in Western Australia (WA) since the 1970s. The DLB is the longest running linkage centre in Australia. The unit was established in 1995 within the Department of Health and has provided record linkage services for the WA population. The linkage system contains over 40 million records for four million people, and has provided data to over 700 research studies.

Data Linkage Unit, Australian Institute of Health and Welfare (AIHW). This record linkage unit undertakes data linkage to support whole-of-government policy development at the Commonwealth level. AIHW provides record linkage services relating to the National Death Index and other national data collections. It works closely with the Data Integration Service Centre in order to support data custodians and researchers in undertaking data integration projects.

SA-NT DataLink. The SA-NT DataLink unit was established in 2009 as part of a collaboration between partners in the Northern Territory (NT) and South Australia (SA). The joint venture includes the governments of SA and NT, and the university sector in SA. SA-NT DataLink is currently involved in a number of demonstration projects dealing with cancer treatment and early childhood development.

United Kingdom (UK)

General Practice Research Datalink (GPRD) [45]. The GPRD is a primary care database containing information from electronic medical records for approximately 8 % of the population of the UK. Over 600 practices contribute data to the GPRD, and the system includes information for over 11 million persons. GPRD data has been linked to national death and hospitalization data, disease registers at the person level, and to socioeconomic and census data at the small area level.

Oxford Record Linkage Study (ORLS) [1, 17]. The Oxford Record Linkage Study began as a joint project between the National Health Service (Oxford Regional Health Authority) and academics (University of Oxford) and was funded largely by the NHS. The Oxford group performs English national linkage with funding from the National Institute for Health Research and continues to use the Oxford subset for ORLS. The system is regarded as one of the earliest sites for record linkage.

Secure Anonymised Information Linkage (SAIL) databank [16]. Established in 2006, SAIL is an anonymous record linkage system that securely brings together a wide array of routinely-collected data for research, development and evaluation. The SAIL databank contains over 500 million records covering the entire population of Wales.

(continued)

Scotland [11, 15, 29]. The Scottish system contains linked hospital, mental health, cancer registration and death records. The data it holds ranges from 1981 to present day and is updated on a monthly basis for the entire Scottish population.

Canada

Manitoba Centre for Health Policy (MCHP) [37]. Providing record linkage services for Manitoba since 1970, this centre has over 100 linked data collections, including datasets from the education and justice sectors. The Population Health Research Data Repository housed at MCHP was developed to describe and explain patterns of health care and profiles of health and illness, facilitating cross-sectoral research in areas such as health care, education and social services. MCHP acts as a steward of the information in the Repository for agencies such as Manitoba Health and provides access to the data for a wide variety of research purposes. Most of the datasets within the Repository are updated annually.

Population Data BC (PopData). Established in 1996 this centre provides record linkage services to British Columbia, Canada. PopData holds individual-level, de-identified, longitudinal data on British Columbia's 4.6 million residents. It provides researchers with access to one of the world's largest collections of healthcare, health services and population health data. These data are linkable to each other and to external datasets, where approved by the data provider. Linkage of data across sectors such as health, education, early childhood development, workplace and the environment, facilitates advances in understanding the complex interplay of influences on human health, well-being and development. Such research informs health related policy-making and investment decisions for healthier communities.

Over time, a large number of record linkage centres have adapted their systems to include a repository of links between major or 'core' datasets. These links are updated on a regular basis and then used by researchers following an application and approval process which addresses ethical, privacy, confidentiality and stakeholder/custodian requirements.

11.1.2 Privacy Challenges in Health Record Linkage

Data from health records are highly confidential, often containing sensitive information. A critical issue in the conduct of health record linkage is striking a balance between using individual medical records for important public good activities and ensuring that such private information is kept confidential [22]. For a record linkage infrastructure to be put in place, data custodians, researchers and record linkage

centres have had to work together to develop data access and usage models that provide sufficient guards to privacy. Record linkage centres, in particular, have developed and implemented various strategies to minimize the risk to privacy posed by their operations. These strategies and techniques can be grouped into three categories:

Data governance regimes: Legal frameworks, policies and procedures around data access, data transfer and IT security.

Operational models and data flows: Organizing and controlling the movement of information to minimize information disclosure risk.

Privacy-preserving linkage techniques: Developing and testing methods of accurate data matching without requiring personally identifying information.

In this chapter, we discuss these three categories of privacy-preserving techniques in more detail, with particular focus on how they have been implemented by linkage centres that are currently in operation.

11.2 Data Governance

The governance around record linkage covers all aspects of the infrastructure including people, processes and information technology. Understanding and using appropriate standards to address legislative and operational requirements, ensures a systematic and safe approach to linkage.

The process of developing a governance framework around linkage services often helps to identify key data controls and to develop the relevant principles and practices required for undertaking record linkage [30]. Areas of responsibility and accountability, typically identified during the design of a data governance framework for record linkage include:

- Legal obligations
- Public interest
- Privacy
- Information security
- Information governance
- Consent
- Anonymization
- Data access and linkage models
- Authorising/advisory bodies
- Data custodians/data controllers
- Cross-sectoral/cross-jurisdictional data sharing
- Consumer (public and stakeholder) engagement
- Sanctions

Some of these areas of governance are specific to individual research projects and require a customised approach to meet the relevant principles. In operational centres,

which facilitate linkage on a regular basis, a key consideration is the *principle of proportionality*. This involves addressing the key requirements identified by the decision makers to achieve optimal governance, which can then be applied within a best practice framework to ensure a safe, effective and proportionate governance model. Some of the requirements can be less flexible than others, e.g., legal obligations, and it is possible that some of the requirements may conflict, necessitating careful consideration to enable linkage. However, the overriding principle is often that the research project is of benefit to society, allowing stakeholders to find a balance between privacy and public good.

11.2.1 Legal Obligations

It is critical that operational linkage centres ensure that their collection, use and disclosure of personal information comply with applicable information privacy laws. Although the regulations and legal requirements vary from one jurisdiction to another, the main drivers are consistent and focus on protecting an individual's privacy and maintaining confidentiality.

Compliance with legal requirements relating to privacy is essential but it is only one dimension of good governance. Equally important is the development of a strong culture of understanding and support for privacy goals and governance best practice. The Pledge of Privacy by the Centre for Health Policy in Manitoba and its associated Privacy Code [33] provides a good example of this.

11.2.2 Information Governance

To maximize privacy and maintain the confidentiality, integrity and availability of the personal information, linkage centres have developed robust information governance frameworks which:

- Provide a consistent and measured approach to good governance and demonstrate that stakeholders understand, prioritize and manage risks associated with data transmission, access, use, storage and disposal;
- Recognise the need for flexibility given the different organizational environments and business requirements of all stakeholders; and
- Take into account the existing information management and security policies, practices, processes and infrastructure of all stakeholders.

Information governance procedures typically contain specific provisions around security, operations, risk management and privacy. These have been incorporated into the approvals process, operational models and linkage methods of many linkage and research centres to ensure a coherent and comprehensive framework which addresses privacy and governance requirements [5, 23, 44].

In Wales, for example, the SAIL databank has developed comprehensive information governance procedures to ensure compliance with data protection legislation and confidentiality guidelines within the NHS Information Governance framework [16]. This framework is based upon the *Data Protection Act 1998*, which concerns data which allows the identification of living persons. To build the SAIL databank, a set of information governance objectives was identified that addressed issues ranging from data transportation, anonymous record linkage, disclosure risk, data access controls and research project approval. Policy documentation and standard operating procedures were developed on the basis of these objectives. An independent audit was subsequently undertaken to evaluate this documentation against legislation and to confirm its use in practice.

11.2.3 Separation of Data and Functions

The privacy risks associated with record linkage have also been addressed through the *separation principle* [28]. Under this principle the process of record linkage, as well as the data items used in a linkage activity, are kept separate from the processes that extract and deliver content or clinical data for researchers.

Although separation of data and functions has been implemented by many linkage centres to maximize the protection of privacy, the additional data flows required under this model also introduce some operational challenges. One obvious challenge is the coordination of numerous ‘separated’ elements before different datasets can be joined up. This process can be complex and requires careful design and implementation in order to avoid bottlenecks in the system.

11.2.4 Application and Approval Process

Privacy risks are also mitigated through the application and approval process. Each research project requires approval from different stakeholders before data can be acquired. Three classes of stakeholders are generally involved:

- *Ethical and Privacy committees*—approval which addresses all organization and jurisdiction requirements (often in coordination with authorising/advisory bodies);
- *Data custodian/data controller*—often requires additional confidentiality and security undertakings for the project;
- *Record Linkage Unit*—includes the implementation of processes and standards for secure management of data throughout linkage operations.

The application and approvals stage requires rigorous and transparent processes that can be applied to each research project. The process often requires binding

agreements related to data release, data confidentiality and security which address specific organizational policies and guidelines. Though important, such processes can be time-consuming and burdensome to those involved [34].

In the Canadian province of British Columbia, a data access committee (DAC) oversees the approvals process [9]. This was instigated to help manage and coordinate the numerous data custodians that would be required to give approval for any single project, and in acknowledgement that each data custodian must consider a research request in the context of other datasets which are simultaneously being requested, and to which their files are being linked. This approach allows custodians to gain a more complete understanding of the privacy risks involved in a project.

The committee ensures that the use of the data is in the public interest, and that the research complies with relevant legislation. Both scientific peer review (to ensure the importance of the proposed study) and ethical approval are required before the final approval by the data access committee.

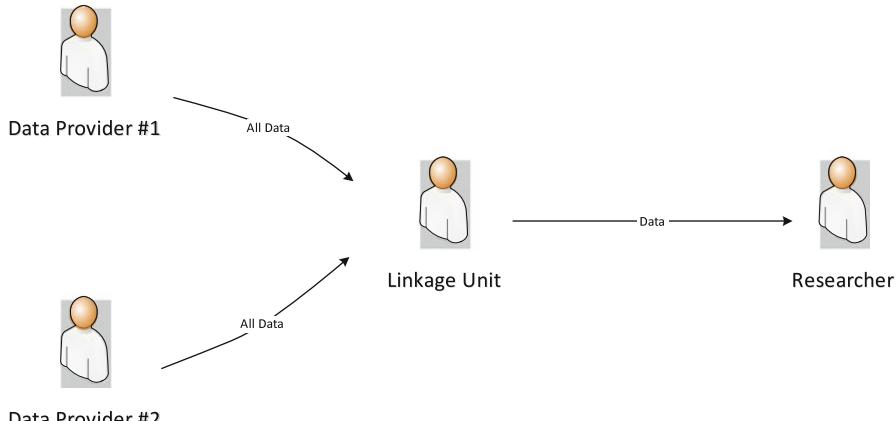
11.2.5 Information Security

Linkage centres operate within a tight information security framework. This determines how the linkage services and infrastructure are to be established and operated in accordance with industry standards and stakeholder requirements. The design of the linkage infrastructure is usually attuned to the relevant, local confidentiality, privacy and security polices to ensure that appropriate measures are in place to sufficiently protect sensitive information.

In Australia, the Centre for Data Linkage designed a secure IT environment in consultation with relevant stakeholders to accommodate datasets from State and Commonwealth organizations, whilst applying the highest level of security [5]. A secure stand-alone network was designed to enable the storage and processing of personal identifiers received from jurisdictional linkage units, researchers and other sources. Several standards and security manuals were used as guidelines for identifying risks and determining appropriate security measures. The environment was subjected to an independent security audit, including a full review of the configuration, operations and use of the infrastructure to ensure compliance with the standards and processes identified by their stakeholders.

11.3 Operational Models and Data Flows

Linkage centres adopt various operational models to ensure that the activities related to record linkage are carried out efficiently and securely. Each model has its own strengths and weaknesses, and their applicability or suitability often depends on institutional settings and governance arrangements. A common theme in all of the



Data Provider #2

Fig. 11.1 A centralized model: data providers give full datasets to the linkage unit, who link and then pass on the data to the researcher

models is the trade-off between simple efficiency (i.e., simple but efficient models with fewer privacy guards) and more complex arrangements (some with intricate data flows that have a greater focus on privacy protection). An outline of some of the more common models is provided below.

11.3.1 *Centralized Model*

In a centralized model, data is supplied to a central linkage facility for matching. The data comprise both personally identifying information and clinical content data. Using personally identifying segments of data, the central linking facility constructs a linkage map that identifies the same individual within and across datasets. A subset of the data is then extracted by the central unit for the researcher (see Fig. 11.1).

In this model there is no separation between the linkage unit and the client services teams who are responsible for liaising with, and provisioning data for the researcher. The model creates a centralized repository of linked clinical data and personal identifiers. Centralized repositories are easier to manage and maintain, and may provide quicker turnaround, as there are fewer opportunities for bottlenecks. Linkage quality may also improve with access to clinical data. The clear disadvantage to this model is that there are no privacy-providing elements inherent in the model—all name-identifying and content information is held by a single organization. The Scottish linkage system is one example of this model [11, 29].

11.3.2 Separated Models

The ‘separation principle’ aims to maximize privacy by ensuring that personally identifying information is kept separate from content or clinical data. This could occur by providing clinical data to one organization and personally identifying data to a separate organization, or it could occur within a single organization by limiting an individual’s access to only one component of the data. The linkage unit requires only the personally identifying components of the data in order to determine which records belong to a single individual, while the researcher only requires clinical information to perform their analysis.

By splitting data in this way, the risk to privacy is dramatically reduced. Even nefarious individuals with access to the clinical data would have difficulty identifying individuals from this information, while those with access to personally identifying information only often have little more information than what could be found in a telephone directory.¹

Several linkage centres across Australia and in Canada and Wales utilize the ‘separation principle’ in their operational models, although implementations vary slightly (see details below) [5, 16, 19, 31, 38, 41].

11.3.2.1 Separated Model, with Centralized Clinical Data Repository

In this model only the personally identifying information required for linkage is supplied to the linkage unit. Specifically, the clinical data is passed to a client services team who are responsible for liaising with, and providing required data to researchers. It is important here that the linkage team is separated from the client services team in the case where they are both part of the same organization. Once the linkage is performed on the personally identifying information, the linkage map is passed to the client services team who join this to the clinical information to create datasets for research and analysis. The client services team subsequently extracts the information for the researcher (see Fig. 11.2).

By adhering to the separation principle, privacy standards are increased and the risk of inadvertent or malicious release of sensitive information is reduced, as no extra individuals have access to both parts of the dataset. This model contains both a centralized repository of person identifiers, maintained by the linkage unit, and a centralized repository of clinical information—this reduces complexity and ensures systems are easier to manage and maintain. In situations where the linkage unit and the client services team are within the one organization, ensuring complete separation may be difficult, and there may be perceived risk that individuals have access to both portions of data. The Western Australian Data Linkage Branch is an example of this operational model [21].

¹However, for certain datasets (e.g., mental health or cancer registries), the existence of an individual’s name within the collection can itself reveal information about that individual.

11.3.2.2 Separated Model, with No Centralized Data Repository

Similarly to the previous operational model, the linkage unit only receives the personally identifying information required for research. However in this model, the clinical data is not sent to a centralized data repository; rather, it is kept by the data providers. Once the linkage map is created by the linkage unit, it is distributed to all data providers. Subsequently, each individual data provider extracts their clinical data for the researcher. The researcher is responsible for amalgamating/merging these extracts into a final linked research dataset (see Fig. 11.3). Linkage centres in South Australia, New South Wales and the CDL in Western Australia (which undertakes cross-jurisdictional record linkage in Australia) have implemented this form of separated model [5, 19, 31].

Without a centralised clinical data repository, there are fewer individuals with access to sensitive data, thereby reducing privacy risks. However, this model requires data providers to play an active role in linkage operations.

While the data flows are relatively simple for a research project involving a single data provider, these data flows quickly become complex for projects involving multiple data providers. For example, for a project with ten data providers, the linkage unit must send the correct portion of the linkage map back to all ten data providers, and all ten data providers must then send their clinical data to the researcher. With more complex data flows, there are more opportunities for bottlenecks and errors in the process. Researchers must also have the expertise required to amalgamate the myriad of datasets they receive from data providers.

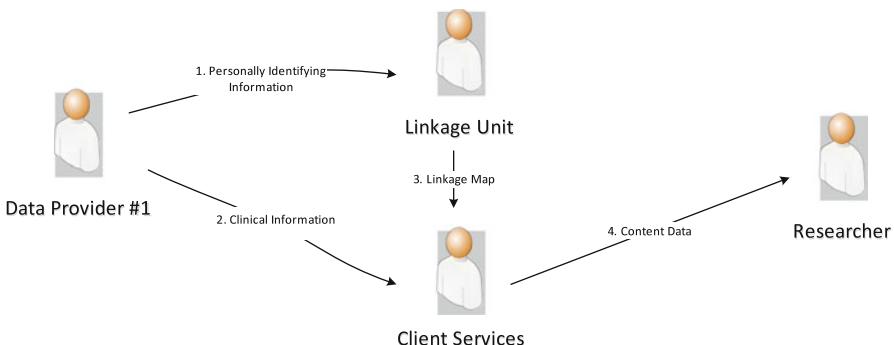


Fig. 11.2 The data provider splits the data, sending the personal identifiers to the linkage unit and the clinical content to the client services team. The linkage unit then provides the linkage map to the client services team who join it to content data to create datasets for research and analysis

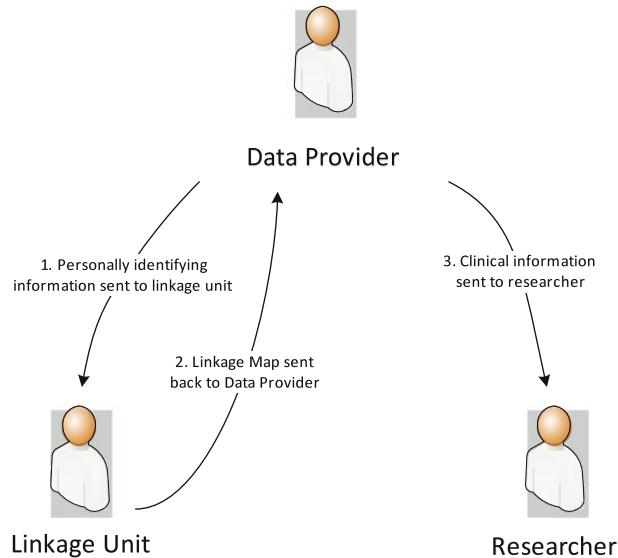


Fig. 11.3 In the absence of a repository of clinical data, this is supplied to the researcher by the data provider

11.3.3 A Technique to Avoid Data Collusion

An additional measure used by most linkage centres to maximise privacy is the use of project specific identifiers. For every research project, the actual identifier which represents a specific person is changed. ‘Person 2’ in one research project is thus a different person to ‘Person 2’ in another study. This reduces the risk of data pooling.

11.4 Privacy Preserving Methods

To lower the privacy risks further, some linkage centres have incorporated various privacy-preserving techniques into their data matching processes. These techniques operate on encrypted information and do not require data custodians to release personal identifiers to third parties. A challenge in the adoption of privacy-preserving methods is to achieve or maintain high linkage accuracy, while information content is degraded. Privacy options for linkage including *Statistical Linkage Keys* (SLK), *phonetic encoding*, *hash encoding* and *bloom filters*. These approaches differ in their methods, maturity, practicality and suitability for large-scale linkage operations.

11.4.1 Privacy Preserving Models

Privacy-preserving techniques can be classified into two general categories—those that utilize a third party for performing the linkage (three party protocols) and those that do not (two party protocols). Two-party protocols typically require a greater amount of necessary communication and computation [39] to compare records, but do not rely on the existence of a trusted third party [42]. In the medical record linkage context, three party protocols are more common.

In a typical three party protocol, data providers must carry out the initial encryption themselves. The data providers need to communicate a shared secret, which is not shared with the linkage unit. The encrypted data is subsequently sent to the third party linkage unit. The linkage unit then identifies which encrypted records belong to the same individual (see Fig. 11.4).

Privacy-preserving techniques generally adopt the same security model as unencrypted linkage; however, there may be differences in the particular privacy algorithm used. Nearly all privacy-preserving protocols adopt an *honest-but-curious* threat model [42], whereby parties are expected to try to carry out the protocol correctly, but will also try and find out as much as they can from any data received.

Privacy-preserving protocols range in the comparison techniques applied—from those carrying out an exact match of entire records to those employing string similarity measures on individual fields. Protocols utilizing more fine-grained techniques in determining similarity will generally yield higher linkage quality [39].

11.4.2 Techniques for Privacy Preserving Linkage

While numerous privacy-preserving methods appear in the record linkage and data mining literature, few of these have been practically evaluated for use in operational record linkage settings [36] and, in reality, very few privacy-preserving methods are used routinely by operational linkage centres. Those methods that have been adopted by operational linkage units have tended to be underscored by a similar philosophy around data flows, namely, *minimum linkage information* (MLI). Most operational centres simply employ methods (or collections of methods) that implement this philosophy in different ways.

11.4.2.1 Minimum Linkage Information (MLI)

MLI methods generally conduct exact matching on a pre-processed subset of personal identifiers. These methods aim to hit a “sweet spot” in the matching process where there is enough information to distinguish all individuals, but a reduced amount of information such that sources of difference between records belonging to the same person are removed (see Fig. 11.5).

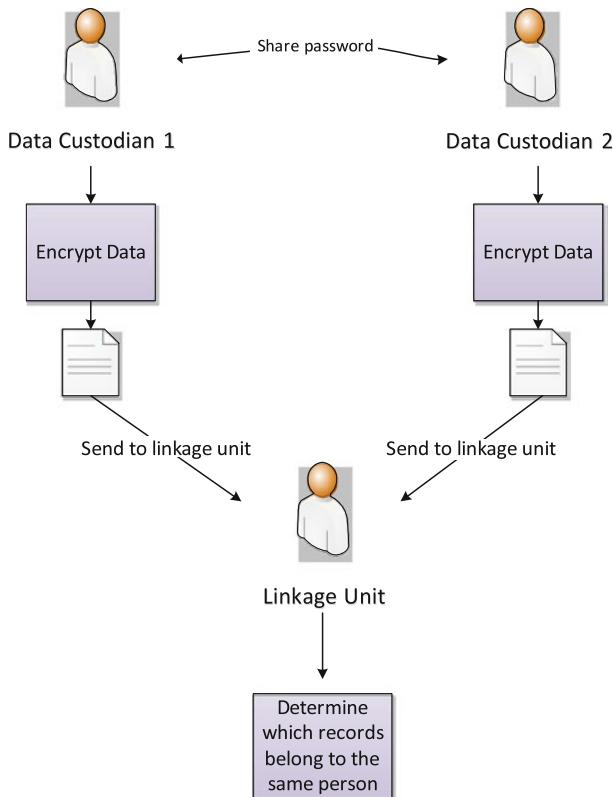


Fig. 11.4 Basic data flow of three party privacy preserving linkage

The information used for exact matching can be easily concatenated and hashed, allowing privacy-preserving linkage to occur. Hashing essentially involves using an algorithm to map an input into a fixed length output. Hash algorithms have several important properties:

- A hash of the same input will always produce the same output;
- Knowing the output, it is not possible to convert back to the input—a hash is one way only; and
- Slight changes in input will completely change the output.

There are various implementations of the MLI approach. The Australian Institute of Health and Welfare (AIHW), for example, uses the second, third and fifth letters of surname, the second and third letters of forename, the full date of birth and the person's gender to create a *statistical linkage key* (SLK), which is used to match records. These characters are simply amalgamated into a single string (the SLK), as shown in Fig. 11.6. Those records containing an identical SLK are said to belong to the same individual. AIHW has used the SLK successfully for a large number of

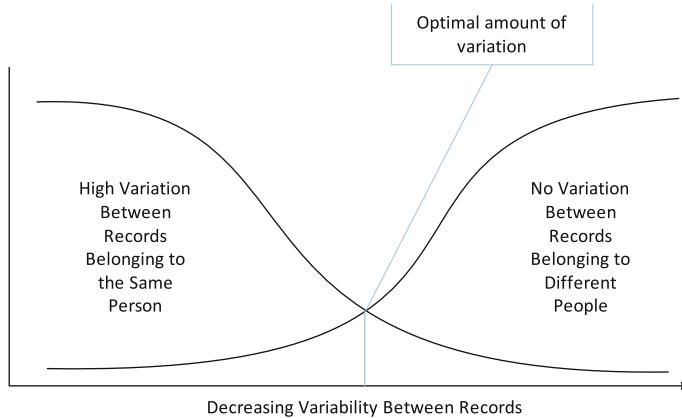


Fig. 11.5 Numerous protocols attempt to reduce the variability between records of the same person, while maintaining variability between records belonging to different people

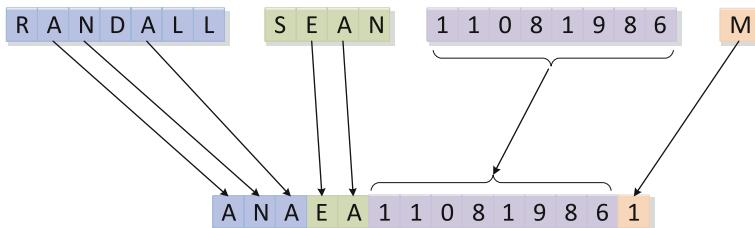


Fig. 11.6 Creating a statistical linkage key

linkages and research projects [25–27]. As the SLK is not hashed, linkages can be performed using rules (deterministic methods) and, in practice, information outside the SLK is used if available.

Similar approaches have been used elsewhere. The Swiss Anonymous Linkage Code [4], for instance, creates a hash from the phonetic codes of first and last name (in the example in Fig. 11.7, these are ‘R543’ and ‘S500’), along with full date of birth and sex. An analogous method has been used to conduct linkage in France [35].

A method similar to both the SLK and the Swiss Anonymous Linkage Code has been proposed by Weber [43]. This method involves taking the first two letters of the first name and the surname, along with the date of birth and gender, and creating a hash of this combination to match on. A similar privacy preserving protocol has been incorporated into the GRHANITE™ system [6]. However, this health informatics system also uses a number of pre-processing steps, including phonetic encoding and nickname resolution, before creating a hash to match on.

A common feature of all of these methods is that they do not use string similarity measures; hence, linkage quality tends to be lower than that achieved through probabilistic methods [3]. However, linkage undertaken using MLI methods can be faster than probabilistic linkage on unencrypted personal identifiers.

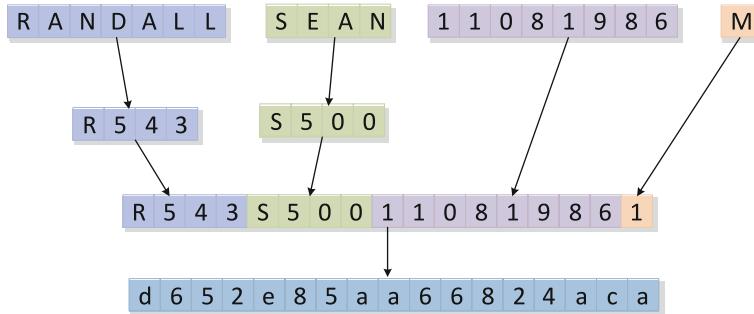


Fig. 11.7 First and last name are phonetically encoded and concatenated with date of birth and sex, which is then hashed to form the Swiss Anonymous Linkage Code

11.4.3 Requirements of a Privacy Preserving Linkage Technique for Operational Linkage Centres

A large number of novel privacy-preserving record linkage (PPRL) techniques have appeared in the literature in more recent times [42]. These protocols differ in their method of preserving privacy, scalability, error tolerance and security. However, very few of these techniques have been practically evaluated for use in operational record linkage settings [36]. For PPRL techniques to be viable options in an operational context, they must not only be secure, but also accurate and efficient.

11.4.3.1 Measuring and Maintaining Linkage Quality

The traditional quality assessment mechanism for record linkage is to manually inspect individual pairs of records in order to identify possible false positive errors. As all records in the privacy-preserving linkage are encrypted, manual inspection of results is no longer possible. Thus, poor quality links cannot be easily identified by a linkage centre, which may compromise the analysis of linked data and potentially lead to erroneous research findings.

In addition to manual inspection for quality assessment (QA) purposes, some linkage units use clerical review as part of an overall quality improvement program, where a small proportion of links are routinely scanned and manually reviewed [17, 20, 31]. While this method results in a high quality linkage, there are drawbacks; it is very expensive and time consuming, and may not be feasible for large linkage projects. Data custodians may also feel uncomfortable about the increased privacy risk when business processes require personal identifiers to be regularly manually examined. Clerical review processes are not used by all linkage units, however; and there is little published evidence of the extent of quality improvement provided by clerical review. This clerical review process, too, is no longer possible when using privacy-preserving linkage techniques.

Manual inspection of record pairs is also used to determine optimal threshold settings in probabilistic record linkage. There is currently no obvious way to determine the optimal threshold with probabilistic encrypted linkage. Some research suggests that the optimal threshold for encrypted linkage appears to be very similar to that of unencrypted linkage, which both have a reasonably wide tolerance (a threshold setting 1 or 2 points either side will give almost equal results) [36]. A rules-based, deterministic linkage paradigm may also avoid this problem, as this method does not involve setting a matching threshold. Deterministic linkage uses a set of rules to specifically outline which combinations of matching variables will result in a record-pair match [18, 27]. The *matching threshold* for deterministic linkage is the decision of which rules, out of all possible rules, will designate a correct match.

For privacy preserving-linkage techniques to be effective in operational contexts, they must be accompanied with comparable quality assessment methods that can assess (and, ideally, improve upon) the quality of linkage output. Some methods for detecting errors in privacy-preserving record linkage settings have emerged [36]; however, a broader range of tools is required. Moreover, methods for correcting linkage errors (not just detecting them) are also needed.

11.4.3.2 Efficiency

Record linkage is a computationally expensive process. The amount of time required for linkage increases significantly in regard to the number of records that are used as input; if the number of records doubles, the required time for processing will roughly quadruple. The inappropriate setting of a linkage parameter or the use of a software package that is designed only for small datasets, can result in a linkage project which takes months (or even years) to complete [14].

To ensure acceptable run times in record linkage, blocking techniques are often used. These techniques limit comparisons to those records which share a minimum level of identifying information. This is important with large datasets as the potential number of comparisons can be too large to process without the blocking step. For PPRL techniques to be useable in an operational context, they must be efficient. The run time for large-scale linkages using privacy-preserving techniques must be similar to that taken using full demographic information. Techniques which utilize deterministic exact-matching methods (such as the MLI methods described above) are often fast enough. Other privacy-preserving techniques which utilize string similarity comparisons may require some form of blocking to allow realistic completion times. Current research into private blocking schemes will be important to linkage centres [2, 40], as these developments will allow the units to process data in the same timeframe as traditional methods.

11.4.3.3 Simplicity for Data Providers

Notwithstanding linkage quality and efficiency requirements, methods for privacy-preserving linkage must also be simple for data providers to understand and to use. A data provider's primary focus is to manage the quality and security of their collections; linking data is not always part of their core business. While data providers are often happy for their dataset(s) to be used for linked research, they typically do not have the resources to conduct linkage themselves.

Current models for privacy-preserving linkage propose that data providers encrypt datasets themselves before passing on to linkage centres. Under such models, data providers play a larger and more critical role in the record linkage process. It is vital that they encrypt datasets in the correct manner for even simple errors, such as switching the first and last name fields in order, will be undetectable to the linkage unit, who will only see encrypted data. In practice, tools need to be developed and distributed (with training) to data providers to facilitate the encryption process.

Privacy-preserving methods that involve passing data back and forth between a data custodian and a linkage unit will generally be too complex. A more preferable option is to have custodians passing the encrypted personal identifiers only once during the linkage process.

11.4.3.4 Security

Privacy-preserving protocols differ in the level of privacy protection they provide. Although increasing privacy and security is the objective of PPRL protocols, there is no specific level of requirement in regards to security. All things being equal, a more secure protocol is preferred over one which is less secure, which in turn is preferred over using unencrypted personal identifiers.

Of the privacy-preserving methods so far described, the SLK provides the lowest level of privacy protection. Although not quite as 'visible' as un-encoded personal identifiers, it is still possible for an operator, or malicious individual, to determine whether an individual exists within a database of SLKs.

Other privacy-preserving techniques encrypt data so that those with access cannot learn any information directly from the encrypted values. Although these encrypted values are vulnerable to frequency attacks, the encrypted/hashed protocols are significantly more complex to break, and typically only a small percentage of information can be revealed in such a case.

Another class of PPRL techniques uses hardened cryptographic techniques, which make it very difficult to learn any information about individuals. Unfortunately, such record linkage protocols are currently computationally intensive, making them impractical for all but the smallest applications of record linkage.

11.5 Conclusion

Record linkage centres throughout the world have adopted a multi-faceted approach to the protection of privacy within their operations. They have instituted a range of measures to reduce privacy risks and to protect the confidentiality of data. These include the adoption of strong IT security and governance frameworks, best practice linkage principles and various operational models that seek to control data flows and minimize data disclosure risk.

Operational linkage centres have also implemented a range of privacy-preserving data matching techniques. Although the development and implementation of some of these techniques are still in their infancy, there is great interest in trialling and adapting them in large operational settings. By reducing privacy-related risks associated with record linkage activities, operational centres enhance opportunities for administrative data to be used in medical and health research and for the wider community to benefit from this endeavour.

References

1. Acheson, E., Evans, J.: The Oxford record linkage study: a review of the method with some preliminary results. *Proc. R. Soc. Med.* **57**(4), 269 (1964)
2. Bachteler, T., Reiher, J., Schnell, R.: Similarity filtering with multibit trees for record linkage. Technical Report Working Paper WP-GRLC-2013-02, German Record Linkage Center, Nuremberg (2013)
3. Bass, A., Garfield, C.: Statistical linkage keys: how effective are they? In: Symposium on Health Data Linkage, pp. 40–45, Sydney (2002). Available online at: <http://www.publichealth.gov.au/symposium.html>
4. Borst, F., Allaert, F., Quantin, C.: The Swiss solution for anonymously chaining patient files. *Stud. Health Technol. Inform.* **84**(2), 1239–1241 (2001)
5. Boyd, J., Ferrante, A., O’Keefe, C., Bass, A., Randall, S., Semmens, J.: Data linkage infrastructure for cross-jurisdictional health-related research in Australia. *BMC Health Serv. Res.* **12**(1), 480 (2012)
6. Boyle, D., Rafael, N.: BioGrid Australia and GRHANITETM: privacy-protecting subject matching. *Stud. Health Technol. Inform.* **168**, 24–34 (2011)
7. Brook, E., Rosman, D., Holman, C.: Public good through data linkage: measuring research outputs from the Western Australian data linkage system. *Aust. N. Z. J. Public Health* **32**(1), 19–23 (2008)
8. Brownell, M., Jutte, D.: Administrative data linkage as a tool for child maltreatment research. *Child Abuse Negl.* **37**(1), 120–124 (2013)
9. Chamberlayne, R., Green, B., Barer, M., Hertzman, C., Lawrence, W., Sheps, S.: Creating a population-based linked health database: a new resource for health services research. *Can. J. Public Health* **89**(4), 270–3 (1997)
10. Christen, P.: Data matching. In: Data-Centric Systems and Applications. Springer, Berlin (2012)
11. Clark, D.: The Scottish medical record linkage system: past, present and future. In: Scottish Health Information Programme Conference. St. Andrews (2009)
12. Ferrante, A.: Developing an offender-based tracking system: the Western Australia INOIS project. *Aust. N. Z. J. Criminol.* **26**, 232–250 (1993)

13. Ferrante, A.: The use of data-linkage methods in criminal justice research: a commentary on progress, problems and future possibilities. *Curr. Issues Crim. Justice* **20**(3), 1–15 (2009)
14. Ferrante, A., Boyd, J.: A transparent and transportable methodology for evaluating data linkage software. *J. Biomed. Inform.* **45**(1), 165–172 (2012)
15. Fleming, M., Kirby, B., Penny, K.: Record linkage in Scotland and its applications to health research. *J. Clin. Nurs.* **21**(19–20), 2711–2721 (2012)
16. Ford, D., Jones, K., Verplancke, J., Lyons, R., John, G., Brown, G., Brooks, C., Thompson, S., Bodger, O., Couch, T., Leake, K.: The SAIL databank: building a national architecture for e-health research and evaluation. *BMC Health Serv. Res.* **9**(1), 157 (2009)
17. Gill, L., Goldacre, M.: English national record linkage of hospital episode statistics and death registration records. Technical Report, Unit of Health-Care Epidemiology, Oxford University, Oxford (2003)
18. Gomatam, S., Carter, R., Ariet, M., Mitchell, G.: An empirical comparison of record linkage procedures. *Stat. Med.* **21**(10), 1485–1496 (2002)
19. Harris, J.: Next generation linkage management system. In: Gray, K., Koronios, A. (eds.) Sixth Australasian Workshop on Health Informations and Knowledge Management, vol. 142, pp. 7–12. Australian Computer Society, Sydney (2013)
20. Holman, C., Bass, A., Rouse, I., Hobbs, M.: Population-based linkage of health records in Western Australia: development of a health services research linked database. *Aust. N. Z. J. Public Health* **23**(5), 453–459 (1999)
21. Holman, C., Bass, A., Rosman, D., Smith, M., Semmens, J., Glasson, E., Brook, E., Trutwein, B., Rouse, I., Watson, C., de Clerk, N., Stanley, F.: A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. *Aust. Health Rev.* **32**(4), 766–777 (2008)
22. Holman, C., Meslin, E., Stanley, F.: Privacy protectionism and health information: is there any redress for harms to health? *J. Law Med.* **21**(2), 473–485 (2013)
23. Jones, K., Ford, D., Jones, C., Dsilva, R., Thompson, S., Brooks, C., Heaven, M., Thayer, D., Mc너너, C., Lyons, R.: A case study of the secure anonymous information linkage (SAIL) gateway: a privacy-protecting remote access system for health-related research and evaluation. *J. Biomed. Inform.* **50**, 196–204 (2014)
24. Jutte, D., Roos, L., Brownell, M.: Administrative record linkage as a tool for public health research. *Annu. Rev. Public Health* **32**, 91–108 (2011)
25. Karmel, R.: Linking hospital morbidity and residential aged care data. Examining matching due to chance. Technical Report AIHW Cat. No. AGE 40, Australian Institute of Health and Welfare (2004)
26. Karmel, R.: Data linkage protocols using a statistical linkage key. Technical Report Data Linkage Series Number 1, Australian Institute of Health and Welfare (2005)
27. Karmel, R., Anderson, P., Gibson, D., Peut, A., Duckett, S., Wells, Y.: Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study. *BMC Health Serv. Res.* **10**(1), 41 (2010)
28. Kelman, C., Bass, A., Holman, C.: Research use of linked health data - a best practice protocol. *Aust. N. Z. J. Public Health* **26**(3), 251–255 (2002)
29. Kendrick, S., Clarke, J.: The Scottish record linkage system. *Health Bull.* **51**(2), 72 (1993)
30. Laurie, G., Sethi, N.: Towards principles-based approaches to governance of health-related research using personal data. *Eur. J. Risk Regul.* **4**(1), 43 (2013)
31. Lawrence, G., Dinh, I., Taylor, L.: The Centre for Health Record Linkage: a new resource for health services research and evaluation. *Health Inf. Manag. J.* **37**(2), 60–62 (2008)
32. Lyons, R., Hutchings, H., Rodgers, S., Hyatt, M., Demmler, J., Gabbe, B., Brooks, C., Brophy, S., Jones, K., Ford, D., Paranjothy, S., Fone, D., Dunstan, F., Evans, A., Kelly, M., Watkins, W., Maddocks, A., Barnes, P., James-Ellison, M., John, G., Lowe, S.: Development and use of a privacy-protecting total population record linkage system to support observational, interventional, and policy relevant research. *Lancet* **380**(Suppl 3), S6 (2012)

33. Manitoba Centre for Health Policy: Privacy Code (2002). http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departmental_units/mchp/media_room/media/MCHP_privacy_code.pdf (2015). Accessed 04 Sept 2015
34. Mitchell, R., Cameron, C., Rambach, M.: Data linkage for injury surveillance and research in Australia: perils, pitfalls and potential. *Aust. N. Z. J. Public Health* **38**(3), 275–280 (2014)
35. Quantin, C., Bouzelat, H., Allaert, F., Benhamiche, A., Faivre, J., Dusserre, L.: How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure. *Int. J. Med. Inform.* **49**(1), 117–122 (1998)
36. Randall, S., Ferrante, A., Boyd, J., Bauer, J., Semmens, J.: Privacy-preserving record linkage on large real world datasets. *J. Biomed. Inform.* **50**, 205–212 (2014)
37. Roos, L., Nicol, J.: A research registry: uses, development, and accuracy. *J. Clin. Epidemiol.* **52**(1), 39–47 (1999)
38. Roos, L., Brownell, M., Lix, L., Roos, N., Walld, R., MacWilliam, L.: From health research to social research: privacy, methods, approaches. *Soc. Sci. Med.* **66**(1), 117–129 (2008)
39. Schnell, R., Bachteler, T., Reiher, J.: Privacy-preserving record linkage using Bloom filters. *BMC Med. Inform. Decis. Mak.* **9**(41), 1–11 (2009)
40. Steorts, R., Ventura, S., Sadinle, M., Fienberg, S.: A comparison of blocking methods for record linkage. In: Domingo-Ferrer, J. (ed.) *Privacy in Statistical Databases: UNESCO Chair in Data Privacy*. Lecture Notes in Computer Science, vol. 8744, pp. 283–298. Springer, Berlin (2014)
41. Trutwein, B., Holman, C., Rosman, D.: Health data linkage conserves privacy in a research-rich environment. *Ann. Epidemiol.* **16**(4), 279–280 (2006)
42. Vatsalan, D., Christen, P., Verykios, V.: A taxonomy of privacy-preserving record linkage techniques. *Inf. Syst.* **38**(6), 946–969 (2013)
43. Weber, S., Lowe, H., Das, A., Ferris, T.: A simple heuristic for blindfolded record linkage. *J. Am. Med. Inform. Assoc.* **19**(e1), e157–e161 (2012)
44. Weisbaum, K., Slaughter, P., Collins, P.: A voluntary privacy standard for health services and policy research: legal, ethical and social policy issues in the Canadian context. *Health Law Rev.* **14**(1), 42–46 (2005)
45. Williams, T., Staa, T.V., Puri, S., Eaton, S.: Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Ther. Adv. Drug Saf.* **3**(2), 89–99 (2012)

Chapter 12

Privacy Considerations for Health Information Exchanges

Dalvin Hill, Joseph Walker, and John Hale

Abstract Health Information Exchanges (HIEs) constitute a powerful mechanism for sharing Electronic Health Records (EHRs) across organizational boundaries in healthcare systems. The electronic sphere of patient data is growing but some patients and medical providers remain hesitant to adopt networked information technology due to privacy and security concerns. The U.S. Government has recognized the importance of safeguarding and preserving the privacy of patient data in HIEs, establishing and endorsing privacy standards and information sharing guidelines. This chapter explores the issues and principles shaping HIE privacy solutions, and discusses emerging trends that will influence the design and implementation of privacy-preserving technologies for HIEs.

12.1 Introduction

The potential benefits of Health Information Exchanges (HIEs) are well understood, and have been described by the Office of the National Coordinator for Health Information Technology (ONC), a part of the U.S. Department of Health and Human Services (HHS), in recent publications [59, 64]. Through streamlined information sharing and interoperable workflow processes across healthcare systems, HIEs enable clinicians to deliver a higher quality of care with greater safety for patients [53]. They can help clinicians reduce medical errors and avoid redundant testing. Along the way, HIEs yield new efficiencies and better clinician access to medical diagnostic tools and clinical decision support systems. They serve as a vehicle for public and community health reporting and monitoring, and offer a valuable link between research and clinical practice.

D. Hill • J. Hale (✉)

Tandy School of Computer Science, The University of Tulsa, Tulsa, OK, USA
e-mail: dalvin-hill@utulsa.edu; john-hale@utulsa.edu

J. Walker

MyHealth Access Network, Tulsa, OK, USA
e-mail: joe.walker@myhealthaccess.net

However, the same features and functions of HIEs conferring these benefits can also pose serious threats to patients' privacy. By participating in HIEs, individuals may be exposed to greater risk of misuse and unwanted sharing of their personal medical information. Increased accessibility inevitably increases the potential for medical identity theft and medical financial fraud, although many steps are taken to address these risks. Controls and safeguards must be thoughtfully architected and deployed in order to mitigate these threats and to provide adequate assurance to individuals that their privacy will be protected [63].

This chapter presents the history and environmental context of privacy in HIEs. Section 12.2 describes the components of a HIE, the foundational architectural models of HIEs, and the prevailing security and privacy governance. Section 12.3 discusses the core privacy issues associated with HIEs. Then, Sect. 12.4 presents principles that HIE stakeholders and implementers alike can adopt as a foundation for preserving patients' privacy. Finally, Sect. 12.5 identifies emerging trends that will inevitably influence the national dialogue on HIE privacy.

12.2 Health Information Exchanges

HIEs constitute an interconnected medium in which physicians, laboratories, pharmacies, and other providers can access and store information about a patient. Information sharing is valuable when rendering diagnosis and treatment, as it offers a more complete view of a patient's history. Additionally, HIEs can enable clinicians to reduce the number of laboratory tests, and accordingly the costs associated with redundant tests. More importantly, they allow clinicians to reduce errors that can materialize during verbal transmission of patient information between clinicians, and/or during the transcription process from a faxed copy. In the U.S., the ONC has supported the development of HIEs in keeping with its mission to field an advanced infrastructure for a national health information system.

A HIE is a complex network of many actors and systems. Its architecture, which may be federated, centralized or hybrid, is designed to permit broad accessibility spanning healthcare providers, functionaries and other constituents. A HIE bridges disparate information systems, relying on common data formats and standards to support the proper interchange of patient information. Even more substantial challenges arise for HIEs at the organizational level, where participation and cooperation are predicated in part on trust. At the level of the individual, privacy and security are critical elements to engender trust [83].

12.2.1 HIE Actors and Systems

Numerous actors in a healthcare system work together to diagnose, treat and care for patients. In the process, they produce and share volumes of personal health

information. At the center of the healthcare system is the patient, who triggers the collection of data when he/she visits a doctor, hospital, or clinic. Additional information can be collected based on a prognosis, and can result in a referral to see a specialist, or to obtain laboratory work. Each provider contributes different information to the system. These actors include clinicians, hospitals, laboratories, pharmacies, allied health providers and insurance companies.

Patients are responsible for supplying clinicians with information about their symptoms, family medical history, and previous medical care. Clinicians and hospitals can use this information to diagnose illnesses and recommend treatment to the patient. Based on the information provided by the patient, clinicians can inform them about current conditions, refer the patient to seek additional tests at a laboratory, or order medication to be picked up at a pharmacy.

Clinicians share information with pharmacies regarding prescription needs, laboratories for testing purposes, locations to perform X-rays and other imaging, and make referrals to other doctors and specialists. Clinicians can also benefit from information shared by pharmacies, results from laboratories, specialists' findings, historical, and current information from other physicians about a patient. Pharmacies accept prescription orders from clinicians to deliver medication to patients.

Clinicians rely on the physical examination of a patient in addition to testing at internal and external laboratories. The process for external laboratories may entail a phone call to a lab to secure an appointment, along with a hand-written prescription ordering a specific type of test(s) to be conducted. Once results are available, they are transported back to the clinician. Laboratory Information Management Systems (LIMSS) are replacing this error-prone and less effective process of manually ordering and transporting laboratory tests and results [66]. Specifically, LIMSS allow labs to capture, store, and transmit lab data in an electronic format, as well as operate instruments and conduct analytical tasks.

Hospital Information Systems (HISs) manage a hospital's administrative, financial and medical information [67]. Such systems capture, record, and store information pertaining to a patient's visit. Prior to HISs, different departments within a hospital stored their information locally and independently. Some departments stored their information in paper-based formats, while others stored their record electronically. To transport information between departments, intra-office mail, a messenger, or fax was used. HISs on the other hand, are used to capture clinical data in an electronic format, and have the capability to share this information with other systems. These systems typically operate within a client/server environment, and the information captured is stored in a database over the network.

Hospitals require a massive amount of storage space for the volume of data that is captured, processed and stored on a daily basis. Information must be captured for billing, appointment, and medical documentation purposes. Medical documentation becomes a part of the patient's record in the EHR [45]. HISs can facilitate some types of data mining, since the collected data can be used to reveal patterns and irregularities. Where analytical capabilities are available, clinicians can utilize information about these patterns to administer better quality care.

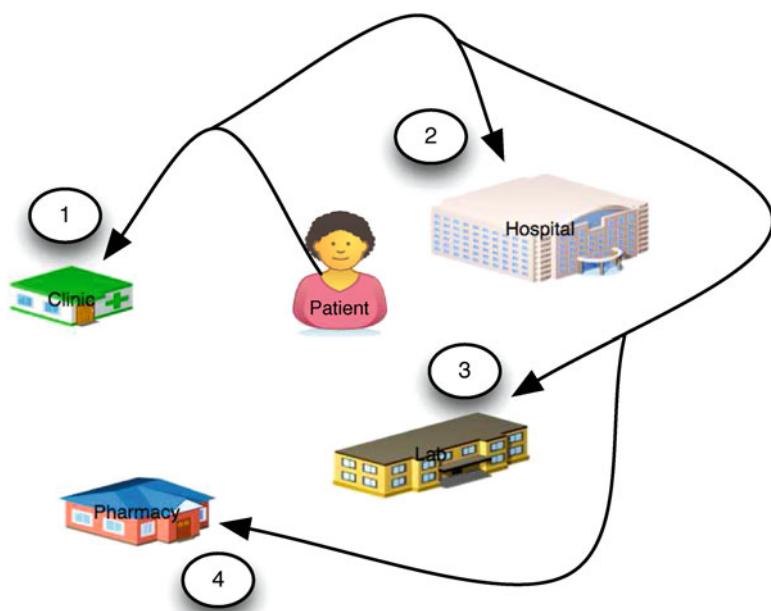


Fig. 12.1 HIE actors and systems

Prescriptions are often ordered via fax, phone, in person with a hand-written record, or sent electronically via a portal. In all instances, excepting electronic submission, the order has to be manually entered into the pharmacy's computer system. This transcription process is prone to errors, as hand-written orders are not always legible and re-entering information is not foolproof. Electronic pharmacy systems can reduce the number of medication errors. These systems can support the pharmacist by automatically conducting safety checks for side-effects, contraindications, and other potential health and safety concerns. It is also easier to maintain a desired inventory level when the computer system keeps track of the quantities of medication on hand. The electronic submission of prescriptions, automation of prescription filling, and labelling systems, are key elements in helping pharmacists reduce the number of pharmacy errors.

Figure 12.1 illustrates a simple use case involving various actors and systems in a healthcare system. In this example, a patient visits a clinic (1), and is subsequently sent to see a specialist at a hospital (2), where testing is done. Test results are evaluated at a lab (3), and a prescription is filled for the patient at a pharmacy (4). Not shown in the figure is the complex exchange of patient information between all these sites, as well as the billing processes and information exchange between all parties and a health plan provider.

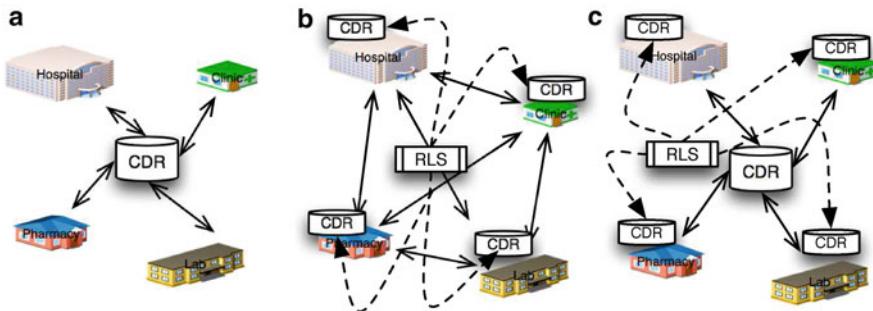


Fig. 12.2 HIE models: (a) centralized, (b) decentralized, and (c) hybrid

12.2.2 HIE Models

The ONC has outlined a roadmap for interoperability of information technology in healthcare systems. The ONC promotes ten principles of interoperability—one of which is to preserve the security and privacy of data in the context of full system connectedness and functionality [59]. This principle is routinely in conflict with many of the other nine. The intention of this inherent tension is to strike the correct balance in a HIE architecture, implementation and practice.

The high-level data architecture of a HIE can be characterized as *centralized*, *decentralized* (federated) or *hybrid*. Each strategy has its own advantages and also encumbers unique disadvantages in terms of functionality, complexity, control and cost in technological and organizational dimensions. Figure 12.2 depicts the centralized, decentralized and hybrid data architectures for HIEs.

In a centralized data architecture [87], all patient data resides in a centralized Clinical Data Repository (CDR). Feeds are established from local participating sites, and data is pushed—either periodically or on-demand—to the CDR. Any access to patient data in the HIE results in queries to a Master Patient Index (MPI), which is used to coordinate extraction of records from a centralized CDR. The centralized HIE model typically requires extensive up-front cooperation and/or well-developed data standardization processes to establish and maintain common data formats and interfaces for integration to the centralized CDR.

In a decentralized or federated model, all patient data is stored, managed and controlled locally [88]. Each site is required to support a common set of functions and services, defined as core to the HIE. Queries to a Master Patient Index are conducted by the Record Locator Service (RLS), which returns paths to patient records stored at local or regional CDRs. Such point-to-point connectivity poses unique challenges in interoperability and often requires heightened technology coordination between HIE participants.

Hybrid HIE models [68], are most common, consisting of some blend of centralized and decentralized elements. Often, some data (commonly conceived as a minimal clinical data set) is stored in a centralized CDR, but a RLS maintains

links to local or regional CDRs for the entirety of data relating to an individual identified in a MPI. Synchronization of data between the centralized CDR and the local/regional CDRs is essential. In a hybrid HIE model, storage and management of patient data is usually under the control of the local participating sites, but access to the patients' data is managed in a more centralized manner.

12.2.3 HIPAA, HITECH and HIE Privacy Governance

The Health Insurance Portability and Accountability Act (HIPAA) was passed by the U.S. Congress in 1996 with five principle objectives: (1) to improve the portability and continuity of health insurance coverage; (2) to combat waste, fraud, and abuse in healthcare systems; (3) to promote medical savings accounts; (4) to improve access to long-term care services; and (5) to simplify health insurance administration [3, 33, 44, 74]. The Privacy Rule in the Administrative Simplification provisions of Title II, establishes conditions to preserve patient privacy. HIPAA also incorporates language intended to ensure the confidentiality, integrity, and availability of patient data in electronic, written or oral form.

With respect to privacy, HIPAA defines the conditions by which an individual's protected health information (PHI) can be collected, used, or disclosed [55–57]. Any individually identifying information that can be traced back to the patient is protected by HIPAA. This includes billing information, data entered by doctors, nurses and other providers, patient information stored in the computer system, and the health insurer's information about treatment and care.

The Privacy Rule allows only authorized users to view PHI when providing treatment to a patient, to pay medical providers for treatment rendered, to perform necessary administrative functions (healthcare operations), and to give mandatory reports to police. Information cannot be legally issued to any other party, except to law enforcement and for public health purposes only, unless a patient gives written permission. HIPAA governs any organization that handles PHI—healthcare providers, purveyors of health plans, and business associates alike (Fig. 12.3).

Following HIPAA, the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 endowed HHS with the authority to, among other things, create programs and establish implementation guidance for private and secure electronic health information exchange [32]. The Omnibus Rule, developed by the Office for Civil Rights, clarified and expanded the scope of HITECH's and HIPAA's privacy and security provisions. It increases the liability of business associates, expands the right of individuals to PHI access and notice, and increases privacy protection for genetic information [20].

12.3 Privacy Issues with HIEs

The sensitive nature of medical information mandates that HIEs pay particular attention to concerns over patient privacy. Most HIEs are built primarily for solving the business problem of getting medical data where it needs to go, while satisfying requirements for privacy and security. While patients are always the final consumers of the healthcare services they receive, most HIEs are designed for the health care providers to be the primary customers.

HIEs are also a newcomer to the healthcare industry's infrastructure, and in order to succeed in establishing their place permanently in the infrastructure, they must often hone their focus to solving one problem at a time. Thus, the features and functions that patients might want, while HIEs often are open to discussing the possibilities, are often relegated to the technology wish list, at least in today's climate.

The ONC has been particularly concerned about ensuring adequate consideration of patient concerns and has organized a workgroup, known as the *Privacy & Security Tiger Team*, whose role has been to advise ONC policy with respect to patient concerns and interests. Over half of the recommendations of this workgroup have been incorporated into public policy so far, and many other recommendations are under consideration for future incorporation [60].

The guiding values that have emerged from this group resemble closely the debates many HIEs have across the country, as they make decisions about how to address consumer privacy. These core guiding values were listed in the first annual summary of this workgroup [62] and include:

1. The relationship between the patient and his or her healthcare provider is the foundation for trust in health information exchange, particularly with respect to protecting the confidentiality of personal health information.

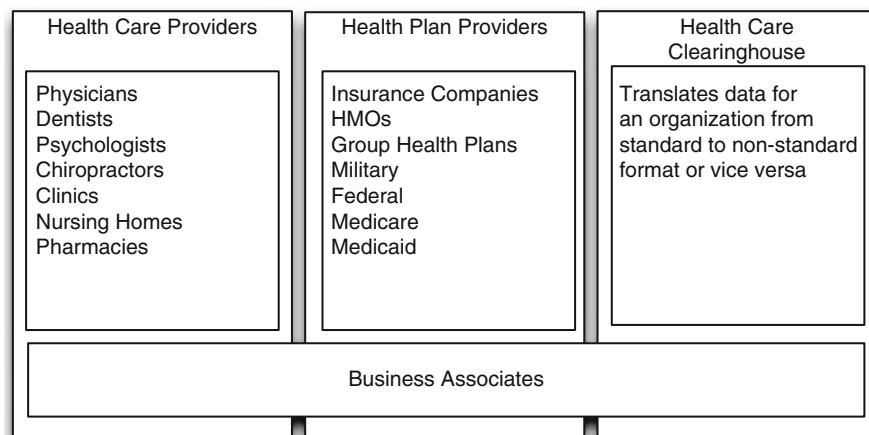


Fig. 12.3 HIPAA covered entities and business associates

2. As key agents of trust for patients, providers are responsible for maintaining the privacy and security of their patients' records.
3. Health IT professionals must consider patient needs and expectations. Patients should not be surprised about or harmed by collections, uses, or disclosures of their personal information.
4. Ultimately, to be successful in the use of health information exchange to improve health and healthcare, the health IT professionals must earn the trust of both consumers and physicians.

12.3.1 Patient Expectations and Concerns

While HIPAA permits healthcare providers to share patient information with one another for treatment, payment and healthcare operations, it also requires providers to be mindful stewards of that information. Patients expect to retain control over their medical information. They expect to know how their information is being used, that it is accurate and that it will remain secure.

Potential misuses of personal health information can cause significant problems for individuals. The severity of the impact from these problems can vary dramatically depending on each individual's circumstances. Victims of domestic violence and popular figures or politicians on the campaign trail, are just a few cases where a medical record in the wrong hands can have catastrophic consequences.

HIEs almost universally acknowledge the need for individuals to be able to limit access to their information by either opting out or by refusing consent for that access. In addition, the *ONC has issued strong guidance to all healthcare professionals that to properly exchange health information electronically, they should ensure to enable “Meaningful Consent for Electronic Health Information Exchange.”* Providing an opportunity for meaningful consent requires considerations in technology, policy, and patient education efforts. This topic is addressed in detail on the HealthIT.gov website [61].

Patients may also rightly worry that their medical information could be used for financial fraud or identity theft. Corrupt or inaccurate medical data could lead to errors in treatment and diagnosis. In addition, every patient-provider relationship is unique, and no person is perfect, or is always perfectly professional. The impact of each record on the first impressions to clinicians for subsequent medical visits is another dynamic that changes with HIE in a way that may surprise patients, for better or for worse. In practice, one finds that there are also many patients who simply do not care how information gets where it needs to go, as long as it causes them as little inconvenience as possible.

Managing and accounting for patient expectations and concerns can be as much of an art as the practice of medicine itself, and no method is going to perfectly address every concern. Many competing interests must be balanced to address patient expectations and concerns effectively.

12.3.2 Tension Between Functionality, Security and Privacy

There is an inherent tension between the operational goals and the security and privacy of any information system. At a minimum, any control implemented to address privacy or security in a HIE adds a layer of complexity. At worst, fundamental design principles and functional requirements may operate in direct conflict with privacy goals. Some of this tension is intentional or unavoidable. In any case, a clear understanding of the dynamics involved is critical for a successful HIE architecture.

Where data collection is concerned, conventional privacy wisdom dictates only acquiring the information needed to perform the task at hand. However, in a clinical setting, the full task (as read upon, for instance, by a differential diagnosis) may not immediately be clear. A more aggressive collection posture could save time and lives. This concept is reflected in the HIPAA Privacy Rule as the *Minimum Necessary* standard [77, 79], which specifically “does not apply to disclosures to or requests by a healthcare provider for treatment” [78].

In terms of use, individuals have legitimate concerns over information gathered for one purpose, being used for another. They may be worried, for instance, that an insurance company may look at elements in their clinician’s EHR, collected for clinical use (diagnosis or treatment), to justify higher premiums. They would also be alarmed if information detailing medical conditions could be sold for marketing purposes (a prohibited practice with regard to PHI in most cases).

Proper privacy protection requires consideration of the conditions for the appropriate disclosure of information [58]. Limited and monitored access to PHI for authorized users is a cornerstone element of any HIE privacy framework. While HIE privacy policies aim to systematically limit disclosure, “break the glass” situations do occur, wherein some safeguards must be overridden to treat a patient in an emergency.

12.3.3 Data Stewardship and Ownership

A HIE privacy framework sets forth rights, roles and responsibilities for actors and agents in the system. Those who store and manage access to PHI are obligated to control the collection, use and disclosure in a manner consistent with the Privacy Rule under HIPAA, and are advised to abide by the privacy principles endorsed by the ONC. Some of the burden of data stewardship shifts away from the end points in a healthcare system to HIE infrastructure providers, particularly when centralization is embraced as an HIE organizing principle. At the other end, decentralized approaches to HIE data architectures may result in varying privacy protection practices, illustrating just how nuanced data stewardship issues can become.

Tangled with data stewardship notions is the core issue of ownership. Simply put, the question is: “*Who owns patient data? the patient, the clinician or the possessor?*” Conceptually, patients often perceive that, as the subject of the data, it belongs to them. However, HIPAA places responsibility for the protection of PHI (and with it, ownership of the records) primarily on the clinicians (“covered entities” under the law) who produce the data. Covered entities often contract with HIEs (“business associates” under the law) and may, in their legal contracts, delegate certain decision-making authority. The law, since HITECH, *holds covered entities and business associates equally accountable for the proper protection of PHI*.

HIEs, if authorized by their data sources, may create new knowledge via analytic tools, potentially adding great value for clinicians and patients, while complicating the question of data ownership. Perspectives of ownership may also change depending on legal agreements and the HIE data architecture, which dictate (legally and practically) storage and management responsibilities. Currently the industry is in flux. Everyone recognizes the patient needs to be more empowered. How to do this, however, is still not clear. Many clinicians and HIEs are, in accordance with new HITECH directives, developing ways for patients to be able to see the data that is stored about them. Some are going further, actually embracing the philosophy that the patient owns the data comprising their medical record.

12.4 Principles and Practice of Privacy for HIEs

This section exposes the fundamental principles of medical information privacy and presents a core collection of privacy-preserving functions and services that are suited for deployment in HIEs.

12.4.1 Guiding Principles

Important roots of modern medical privacy policy in the U.S. can be found in a study conducted by The Secretary’s Advisory Committee on Automated Personal Data Systems within the Department of Health, Education, and Welfare [15]. The study, referred to as the “HEW Report”, established a Code of Fair Information Practices comprising five ideals:

1. **Openness:** There must be no personal-data record-keeping systems whose very existence is secret.
2. **Transparency:** There must be a way for an individual to find out what information about him is in a record and how it is used.

(continued)

3. **Control:** There must be a way for an individual to prevent information about him obtained for one purpose from being used or made available for other purposes without his consent.
4. **Correction:** There must be a way for an individual to correct or amend a record of identifiable information about him/her.
5. **Integrity:** Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take reasonable precautions to prevent misuse of the data [15].

The first two ideals are intended to guarantee that individuals are aware of bodies collecting their private information and how that information is being used. The second two principles give to individuals powers to limit the use and preserve the accuracy of information collected about them. The fifth principle recognizes the responsibility of stewards of personal data to maintain its security. These five principles underpin most modern privacy legislation and regulatory frameworks.

ONC has further developed these ideas into a list of eight principles in the recently-released nationwide interoperability roadmap publication [59]. These eight principles are:

1. **Individual Access:** Simple and timely access for individuals to their personal health information.
2. **Correction:** Ability of an individual to dispute and correct erroneous personal health information.
3. **Openness and Transparency:** Openness and transparency of policies, procedures and practices relating to an individual's medical records.
4. **Individual Choice:** Ability of an individual to make informed decisions regarding the collection, use and disclosure of their personal health information.
5. **Collection, Use and Disclosure Limitation:** Appropriate control of the collection, use and disclosure of personal health information governed by the necessity to accomplish a specific purpose.
6. **Data Quality and Integrity:** Reasonable effort to guarantee personal health information has not been altered inappropriately and that information is accurate.
7. **Safeguards:** Application of security controls to preserve the confidentiality, integrity and availability of medical records and personal health information.

(continued)

8. **Accountability:** Monitoring and reporting of events and actions in HIEs that potentially constitute breaches, misuses or violations of privacy and security [59].

Each of these principles can be traced rather directly to one or more elements of the Code of Fair Information Practices. Individual access is a notable extension that target rights to individuals to view information that is held about them. Patients have the right to timely access of their PHI as held by covered entities, and to a record of PHI disclosures. Accountability recognizes the importance of the security property of non-repudiation in support of privacy-preservation strategies.

These principles balance the criteria that provide allowances for the collection, use and disclosure of PHI without individual authorization or consent. HIE participants are challenged to honor these principles as they seek to solve the business and technical issues to enable the proper flow of data through their systems. Covered entities under HIPAA must ensure they are accurately setting patients' expectations as they develop and distribute notices explaining organizational privacy practices.

12.4.2 HIE Privacy in Practice

The guiding principles promoted by ONC to guarantee the privacy of constituents in a HIE are embodied by a collection of core services. This constellation of controls work in concert as mixture of policy, process and technology to achieve the privacy goals of HIEs. Each fulfills some requisite aspect of functionality to operationalize a subset of the principles described above. They are:

1. **Consent Management:** Informed consent to document collection, processing, storage and distribution.
2. **Access Control:** Policies and enforcement mechanisms that limit the dissemination of PHI.
3. **Data Curation:** Quality control and integrity remediation solutions to ensure the accuracy of patient data.
4. **Secure Transmission and Storage:** Encryption and physical isolation strategies to prevent unauthorized disclosure of PHI.
5. **Privacy Preserving Record Linkage:** Statistical methods over PHI permitting meaningful data analysis, while preserving patients' privacy.
6. **Audit:** Processes to monitor, log and report access to PHI.

Consent Management Nearly all HIEs in practice today offer patients some form of meaningful consent. While HIEs are still struggling to establish themselves in the healthcare industry’s infrastructure, most HIEs employ fairly simplistic patient consent models: either *opt-in* or *opt-out*. These models both comply with HIPAA requirements, though depending on which model they choose, other controls may exist to ensure that only data that is permitted to be shared for permitted purposes to permitted individuals is shared with appropriate permission.

With an opt-in model, patients must learn of the HIE’s existence and explicitly authorize their providers to access the records. Some opt-in models require patients to authorize each potential data source to release records to others via the HIE, while other opt-in models have the data pre-loaded or pre-linked in the HIE, and merely require patient consent for providers to access the data. Some HIEs allow this consent to be given electronically, while others require physical documents to be signed. In some cases providers are required to attest that the patient has authorized the access, and to maintain documentation to back up their claims of authorization.

Some HIEs are required by state or other local laws to operate as opt-in networks. Some HIEs have tailored their consent process to authorize the exchange of sensitive data, such as behavioral health data, which requires specific consent under federal laws [80], in addition to the more common PHI allowed to be shared under HIPAA for permitted purposes. To be successful, a HIE with opt-in must carefully consider the clinician workflow and gain buy-in from its participating clinicians so the consent process is not overly disruptive. Coordinating a consistent change in workflow for many organizations in a HIE is a big challenge, but approximately half of the HIEs operating today have adopted this model.

With an opt-out model, patients are notified about the HIE and their option to opt out either through the participating clinicians’ Notice of Privacy Practices, be explicit “opt-in” like notification in their check-in paperwork, or by some other method. In these HIEs, patient data is generally available to authorized users of the HIE, unless the patient explicitly forbids it. Data that requires explicit consent, such as behavioral health data, is generally not exchanged in these HIEs. Some HIEs utilizing the opt-out model rely on integration with the clinicians’ information systems so that a patient’s medical record is available from the HIE only after the clinician’s information system sends the HIE a medical record for the patient, evidencing that the clinician is in fact treating the patient.

Other HIEs simply require the clinicians to attest that they are only accessing medical records for patients with whom they have a treatment relationship, or for some other permitted purpose under HIPAA. Such models are usually accompanied by appropriate audit controls so an abusive user is likely to be detected.

Regardless of which model is employed, a patient’s consent status is usually “all-in” or “all-out.” In other words, if a patient has expressly opted out of participation in the HIE, it means none of his or her records may be accessed by an authorized HIE user. Conversely, if a patient has expressly opted in, or in the case of an opt-out model, has not opted out, it means that all of their records may be accessed by an authorized HIE user. One common exception to this policy is a “break-the-glass” feature, which some, but not all HIEs support. This feature, where available,

enables HIE users with proper authorization the ability to override patient consent by attesting that an emergency condition exists with this individual, and that such access is necessary to save a life or prevent harm [8, 22]. Such access often triggers additional audit requirements designed to deter and detect any abuse of this access.

In general, HIE administrators recognize that the “all-in” or “all-out” approach to HIE is not the ideal way of handling consent. However, greater granularity of access requires greater technological investment in the infrastructure. Some HIEs are working on advancing the granularity of patient consent, working on or offering the ability for patients to limit what’s in their record by source, or by type of data.

The ONC has tried to help the effort by undertaking the Data Segmentation For Privacy (DS4P) initiative, which sets out a framework for organizing data and choices in a way that consumers can make meaningful decisions. This was further pursued by the U.S. Department of Substance Abuse and Mental Health Services Administration (SAMHSA) to address the fact that mental health facilities are often unable to participate in HIE due to additional federal consent constraints (42 CFR Part 2) that go well beyond HIPAA constraints. SAMHSA’s effort is an open-source project, called Consent2Share [81], which is designed to enable patients to essentially select which HIE users are allowed to see what types and sources of their data. Consent2Share was piloted with a county-level HIE in 2014 [5], to demonstrate that this level of granular consent is possible. The adoption, however, of this or other granular patient consent models is still very new (or non-existent) on a large scale in the industry.

Another challenge with patient consent is helping the patient understand the benefits and risks associated with participating in HIE. While HIEs attempt to explain the risks and benefits (typically in writing), each individual faces different consequences depending on the contents of their medical records, their relationships, and more. Standards-based approaches and architectures provide greater opportunity for systemic interoperability of consent processes; some designs yield solutions in which consent is the driving feature [6, 34, 69]. Recent research has demonstrated potential for developing consent-based systems that can analyze a patient’s record and produce some patient-specific decision support to suggest relevant risks and benefits to opting in or opting out, when a patient is faced with that choice [84]. This level of consumer engagement, however, is far from reality in most of today’s HIEs.

Access Control HIE access control services express policies and implement enforcement mechanisms that dictate permissible access by authorized users to PHI. The keys to effective access control in HIEs are interoperability across federated domains and fine-grained expression of policy. HIPAA requires, and HIEs almost universally employ, some sort of Role Based Access Control (RBAC), and/or Attribute Based Access Control (ABAC). These models require that users be assigned to roles or attributes that correspond with the type of work they perform, so their access is restricted in accordance with the minimum necessary standard of the HIPAA Privacy Rule. These models are fairly straightforward, and solutions that extend, augment or blend RBAC and ABAC exist [2, 89].

Data Curation Data curation implies the ability to deliver patient information and medical records with high integrity and accuracy to users of HIEs. This can be a real challenge when integrating data feeds from multiple organizations, engaging heterogeneous information management platforms and technologies [7, 18, 27, 35, 37]. Frameworks that address data quality as a first order concern pay a price up front for the added investment they encumber, but reap benefits down the road. While technology can play a strong role in providing services that help promote data integrity for a HIE, the most important element is a robust quality assurance process that enforces good data hygiene and standardized formats.

Many HIEs meet this challenge by either displaying exactly what they receive from sources (leaving the user to decide if the information is helpful or not), or by first grouping together data of similar kinds from all sources before displaying them (e.g., all labs from all sources are displayed together, followed by all medications from all sources). The few HIEs who currently go beyond this level of data curation face significant struggles, because of the differences between medical terminology standards used by different medical record systems, inconsistencies in the way similar concepts may be represented by the same medical record systems, and the challenge of successfully translating between the types of data coming in and the universal standard chosen for the community. Companies have arisen specializing in addressing this specific problem, but the difficulty of the work is reflected in the cost of the service, and few HIEs are able to afford these types of services.

HIE frameworks that address data quality as a first order concern pay a price up front for the added investment they encumber, but truly do reap benefits down the road. Effective aggregation and analytics, which can be huge value-drivers for HIE customers, depend on quality data. While technology can play a strong role in providing services that help promote data integrity for a HIE, the most important element is a robust quality assurance process that enforces good data hygiene and standardized formats.

Secure Transmission and Storage Encryption is a core service for securing information during transmission and in storage [17, 73]. It is a logical isolation strategy (physical isolation may accomplish the same end) that scrambles data in such a way that only authorized users may decipher it. The challenge of deploying encryption across any far-flung enterprise is *managing the keys that control encryption and decryption*. Fortunately, standards-based systems and protocols that use encryption to create secure communication channels and storage areas for information networks and systems are common-place and promote interoperability [12, 16, 76]. However, careful thought must be given when deciding on an encryption solution, in terms of its impact to *system function, performance and scalability*. In HIEs, alternative encryption schemes, such as Attribute-Based Encryption (ABE), that seek a more flexible model for logically isolating data in EHRs have been studied [1, 50].

Additionally, encryption must be properly implemented in order to achieve its intended goal, as security experts know well that poorly implemented encryption, even with the greatest algorithms available, affords little protection against knowledgeable attackers. In this age of cyber-crime, HIEs must assume an aggressively

defensive posture, and must expect that they are and always will be targets of aggressive attacks. HIE security must not be disregarded, and HIEs, to properly protect patient data, are required to have a HIPAA IT Security Officer. Though this person may have other roles, the security industry advises this be a single role, as the energy, resources and expertise required to perform this role well can be significant.

Privacy Preserving Record Linkage The fundamental goal of record linkage is to determine when two or more records refer to the same individual [21, 29, 54]. In the context of HIEs, successful linkage is a critical objective for system usability, as well as to ensure that a patient's consent is correctly associated with all of that individual's specific records. For those HIEs that support data analytics, successful linkage is required to accurately represent each individual as a single entity and not as multiple individuals. For patient safety, it is imperative that records belonging to separate individuals not be falsely linked together. While the problem of associating two records to the same person from different sources can seem simple at times (how hard is it to match first name, last name, date of birth, and a phone number), the reality is that in a world where the data sources were never intended to be linked together, where demographics are hand-entered many times by different individuals into different systems with different designs, high-confidence in successful matching with a population of any significant size quickly becomes a real problem.

Privacy preserving record linkage techniques may encompass a range of string comparison, data mining and statistical methods to address the potential for misuse, while permitting analysis [10, 29, 72]. Companies have formed around solving this problem, and open-source solutions have been developed, but algorithms that perform the best always must be tuned to the population they are serving and there is always a need for human eyes to oversee and/or carry out the matching of at least a fair number of records.

Audit In addition to facilitating HIPAA requirements, the audit functionality is an essential component in any HIE's privacy and security framework. Accountability and non-repudiation is achieved in information systems through systematic logging of user access to data. HIE audit services capture and record access requests to PHI [4, 31, 75, 82]. Specific functionality varies based on architecture, but in many ways audit logs and monitoring capabilities can be the primary means of detecting problems in HIE. HIEs wrestle with the same audit challenges as other information systems, including sheer volume of records generated by auditing services, decentralization of audit logs for various types of events, and investment or development of appropriate intelligent and selective audit policies supported by data reduction tools [7, 18, 27, 35, 37].

12.5 Emerging Issues

Information technology has driven advances in medicine, some of which introduce new hazards in security and privacy. This section explores the privacy aspects of technological advances in three domains: Big data, mobile health (“m-health”) and telemedicine, and medical devices.

12.5.1 *Big Data*

As health IT workers advance the capabilities of EHR systems to create much richer patient data sets, new opportunities for clinicians emerge, along with new risks to patient privacy. For this reason, many HIEs and HIE participants resist or avoid architectures that facilitate aggregation and large-scale analysis. However, those who develop frameworks that permit aggregation and analysis do so in a controlled way. These organizations are blazing the trail for “Big data” in healthcare.

Large data sets prospectively offer a much more complete characterization of patient health, condition and state. This can create opportunities for clinicians to identify high-risk patients sooner, potentially enabling clinicians to pursue treatment options that, for example, may prevent heart attacks instead of waiting for them to happen. This also makes possible other positive opportunities for advancing the field of medicine and for improving public health through appropriate and controlled research. New applications like this quite naturally result in an interesting and complex dilemma with respect to maintaining the balance between privacy preservation and appropriate information sharing.

One significant aspect of this dilemma involves preserving each individual’s anonymity within a large, supposedly de-identified dataset. Many factors must be considered to ensure true de-identification. For example, medical image data and full genomes may be included in datasets that previously contained only common elements of an EHR (such as labs and medications). These new data elements pose identification risks that must be taken into account. Medical image data can include visual features that may be used to uniquely identify a subject [48, 49, 86]. Structural MRIs, for example, are of such a resolution that high fidelity facial reconstruction is well within the realm of possibility [71]. It has been shown that such subject identification methods could possibly be performed systematically. Therefore, approaches to protect privacy for medical images may often include strategic removal or transformation of identifying features, which is non-trivial [9, 71].

A similar situation exists for genetic information, where DNA profiling can effectively “fingerprint” a patient [51]. It has been shown that statistical methods may be employed to discern the identity of an individual in a genome-wide association study (GWAS) [70]. In the sense that the data itself is a uniquely identifying and intrinsic feature of a subject, this makes de-identification, without removal of meaningful analysis variables, a particular challenge.

With genomic data, the impact of disclosure of genomic information is potentially much more impactful [11]. Genomic data reveals characteristics not only about an individual, but also about relatives. In addition, genomic data may contain information about future conditions a subject will or may develop. Approaches to preserving privacy in genomic data sets includes partial or limited release of genetic sequences and statistical degradation of the data [52]. However, erosion or elimination of identifying features in genetic sequences (as is the case with images) may remove the very information needed to diagnose or treat a patient.

12.5.2 m-Health and Telemedicine

Mobile, wireless and broadband networks extend the reach of healthcare systems beyond the hospital and clinic, into the home and on the road. They have expanded access to treatment and care for underserved and economically challenged populations, through mobile health applications and telemedicine. Using mobile platforms, patients are increasingly being offered access to their personal health records and in some cases, can consult with clinicians from virtually any location [13, 43, 46, 85]. Some studies have shown that telehealth solutions can reduce costs and enhance the patient experience [14, 23, 36]. Both mobile medicine and telehealth can potentially improve efficiency and quality of care in health systems.

The collection and analysis of location and temporal data on patients afforded by mobile telehealth technology presents new opportunities for clinical insights and improved medical outcomes. Understanding individual behavior patterns as a function of time and space, for example, may yield better strategies for achieving higher adherence rates in treatment regimens. In terms of public health, better tracking of patient activity may reveal enlightening patterns and trends of disease and wellness for an entire community. Residual benefits of leveraging geospatial and temporal information may also include improved security and fraud detection.

Mobile and ubiquitous networking has the potential to transform the role of HIEs in patient care, with more timely and universal access to patient data. Telehealth applications that involve remote patient monitoring and virtual clinician encounters will benefit dramatically from an intelligent integration of HIE services. Seemingly, the rise of HIEs and the adoption of telehealth technologies certainly can go hand-in-hand. However, as such technology allows for patient monitoring on a tremendous scale and if it is allowed to become linked with health information systems, potent questions are raised about potential privacy invasions. It now becomes possible for clinicians (as legitimate users) and abusers alike to create a much more complete picture of the daily life of patients. Besides exposing personal habits, even with de-identified data, the potential exists to re-identify individuals from such data alone.

Another issue concerns the drive to make HIE data available on remote and mobile platforms. From a security perspective, the introduction of mobile devices significantly expands the attack surface of HIEs [28, 47]. Smart phones and tablets

not only embody new points of exposure to an information system, the very nature of mobile technology poses a heightened challenge for security management. Controls and policies must be introduced across the HIE ecosystem as safeguards to mitigate security threats encumbered by mHealth.

12.5.3 Medical Devices

In a different way, patient medical devices may extend the boundaries of HIEs as patients become more direct participants in their own care. Increasingly, medical devices are configured with wired or wireless network capabilities to enable the seamless transfer of data to a patient data repository. Healthcare providers, researchers and technologists have embraced and pursued a vision of patient care in which devices operate in harmony with medical information systems [25].

A collection of standards has been developed to support connectivity and interoperability of medical devices with some clinical information systems housing EHRs [19, 24, 38–42]. Data flowing from a medical device may include treatment regimens, some measures of their efficacy, physiological information, as well as metadata about the patient and the device. These devices have the potential to create a closed loop system in which devices provide a continuous data feed and can be controlled remotely and/or automatically [65]. HIEs will most likely play an integral role in such environments.

The ability to stream and collect data directly from medical devices to systems linked to a HIE can be a powerful informational adjunct for diagnosis and therapy, but also represents a significant threat to patient privacy. Treatment data coming from devices that is decorated with time and location information adds compelling detail to a picture of patient behavior and activity. Such information may be very useful to clinicians, but also to those seeking to invade the privacy of the patients. In most cases, the very fact that a patient is associated with a medical device constitutes PHI that must be adequately protected.

The devices themselves represent a possible point of entry or compromise for an attacker [26, 30]. Devices are often built or bridged with commercial technologies (applications or operating systems) and may be vulnerable to known exploits. An attacker could possibly subvert a device to corrupt an information feed into a medical record system, or perhaps use the device to access information about an individual. In addition to the usual tensions that exist between security and functionality, medical devices, particularly implantable medical devices, are often developed by resource-challenged, market-driven companies, which can constrain the security/privacy solution space for engineers. Thus, a HIE that extends to incorporate data or access from or to such devices must properly account for the potential vulnerability in the medical devices to adequately ensure the privacy of individuals.

12.6 Conclusion

Navigating the privacy issues associated with Health Information Exchanges (HIEs) remains a challenge as information technology and medical applications continue to evolve. Where trust is essential to adoption, striking the right balance between privacy and function can be difficult. A robust regulatory ecosystem steeped in sound principles is one key to success. Another is the establishment and implementation of standards and best practices shaped by organizational realities. As HIEs assume an increasingly large role in the health IT infrastructure supporting the healthcare industry, the pursuit of patient privacy through thoughtful dialogue and strategic action is vital to success.

References

1. Akinyele, J.A., et al.: Self-protecting electronic medical records using attribute-based encryption Cryptology ePrint Archive, Report 2010/565 (2010). Available from <http://eprint.iacr.org/>
2. Alshehri, S., Raj, R.K.: Secure access control for health information sharing systems. In: 2013 IEEE International Conference on Healthcare Informatics (ICHI). IEEE (2013)
3. Annas, G.J.: HIPAA regulations: a new era of medical-record privacy? *N. Engl. J. Med.* **348**(15), 1486–1490 (2003)
4. Appari, A., Johnson, M.E.: Information security and privacy in healthcare: current state of research. *Int. J. Internet Enterp. Manag.* **6**(4), 279–314 (2010)
5. Behavioral Healthcare.: Projects aim to segment data for privacy. <http://www.behavioral.net/article/projects-aim-segment-data-privacy> (2015)
6. Bonnici, C.J., Coles-Kemp, L.: Principled electronic consent management: a preliminary research framework. In: International Conference on Emerging Security Technologies. IEEE (2010)
7. Botsis, T., et al.: Secondary use of EHR: data quality issues and informatics opportunities. In: Proceedings of AMIA Summits on Translational Science, p. 1 (2010)
8. Brucker, A.D., Petritsch, H.: Extending access control models with break-glass. In: Proceedings of the 14th ACM Symposium on Access Control Models and Technologies. ACM (2009)
9. Cao, F., Huang, H.K., Zhou, H.Q.: Medical image security in a HIPAA mandated PACS environment. *Comput. Med. Imaging Graph.* **27**(2), 185–196 (2003)
10. Churches, T., Christen, P.: Some methods for blindfolded record linkage. *BMC Med. Inform. Decis. Mak.* **4**(1), 9 (2004)
11. Claerhout, B., DeMoor, G.J.E.: Privacy protection for clinical and genomic data: the use of privacy-enhancing techniques in medicine. *Int. J. Med. Inform.* **74**(2), 257–265 (2005)
12. Daemen, J., Rijmen, V.: The Design of Rijndael: AES—the Advanced Encryption Standard. Springer, New York (2013)
13. Dégliise, C.L., Suggs, S., Odermatt, P.: SMS for disease control in developing countries: a systematic review of mobile health applications. *J. Telemed. Telecare* **18**(5), 273–281 (2012)
14. DelliFraine, J.L., Dansky, K.H.: Home-based telehealth: a review and meta-analysis. *J. Telemed. Telecare* **14**(2), 62–66 (2008)
15. Department of Health, Education and Welfare.: Records, computers and the rights of citizens: report of the secretary's advisory committee on automated personal data systems (1973)
16. Dierks, T.: The transport layer security (TLS) protocol version 1.2. Internet Engineering Task Force, Networking Group (2008)

17. Diffie, W., Hellman, M.E.: New directions in cryptography. *IEEE Trans. Inf. Theory* **22**(6), 644–654 (1976)
18. Dixon, B.E., McGowan, J.J., Grannis, S.J.: Electronic laboratory data quality and the value of a health information exchange to support public health reporting processes. In: AMIA Annual Symposium Proceedings, vol. 2011. American Medical Informatics Association (2011)
19. European Committee for Standardization (CEN).: Interoperability of patient-connected medical devices (INTERMED) (1997)
20. Federal Register.: 45 CFR Parts 160 and 164 Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; Other Modifications to the HIPAA Rules; Final Rule (2013)
21. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. *J. Am. Stat. Assoc.* **64**(328), 1183–1210 (1969)
22. Ferreira, A., et al.: How to break access control in a controlled manner. In: 19th IEEE International Symposium on Computer-Based Medical Systems (2006)
23. Ghosh, R., Heit, J., Srinivasan, S.: Telehealth at scale: the need for interoperability and analytics. In: Proceedings of the 1st International Workshop on Managing Interoperability and Complexity in Health Systems (MIXHS '11), pp. 63–66 (2011)
24. Glass, M.: ANSI/IEEE 1073: medical information bus (MIB). *Health Informatics J.* **4**(2), 72 (1998)
25. Goldman, J., Schrenker, R., Jackson, J., Whitehead, S.: Plug-and-play in the operating room of the future. *Biomed. Instrum. Technol.* **39**(3), 194–199 (2005)
26. Grimes, S.L.: Medical device security. In: 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEMBS'04, vol. 2 (2004)
27. Gritzalis, D., Lambrinoudakis, C.: A security architecture for interconnecting health information systems. *Int. J. Med. Inform.* **73**(3), 305–309 (2004)
28. Gunter, C.A.: Building a smarter health and wellness future: privacy and security challenges. In: ICTs and the Health Sector: Towards Smarter Health and Wellness Models, OECD Publishing, Paris France pp. 141–157 (2013)
29. Hall, R., Fienberg, S.E.: Privacy-preserving record linkage. In: Privacy in Statistical Databases. Springer, Berlin/Heidelberg (2010)
30. Halperin, D., et al.: Security and privacy for implantable medical devices. *IEEE Pervasive Comput.* **7**(1), 30–39 (2008)
31. Harno, K., et al.: Health information exchange and care integration. *Int. J. Adv. Life Sci.* **1**(1), 46–57 (2009)
32. Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 (2009)
33. Health Insurance Portability and Accountability Act of 1996. Public Law No. 104-191, 110 Stat. 1936 (1996)
34. Heinze, O., et al.: Architecture of a consent management suite and integration into IHE-based regional health information networks. *BMC Med. Inform. Decis. Mak.* **11**(1), 58 (2011)
35. Hovenga, E.J.S., Grain, H.: Clinical decision support systems: data quality management and governance. *Health Inf. Gov. Digit. Environ.* **193**, 362 (2013)
36. Hunkeler, E.M., et al.: Efficacy of nurse telehealth care and peer support in augmenting treatment of depression in primary care. *Arch. Fam. Med.* **9**(8), 700 (2000)
37. Iakovidis, I.: Towards personal health record: current situation, obstacles and trends in implementation of electronic healthcare record in Europe. *Int. J. Med. Inform.* **52**(1), 105–115 (1998)
38. ISO/IEEE 11073-10101.: Health Informatics Point-of-Care Medical Device Communication Part 10101: Nomenclature (2004)
39. ISO/IEEE 11073-10201.: Health Informatics Point-of-Care Medical Device Communication Part 10201: Domain Information Model (2004)
40. ISO/IEEE 11073-20101:2004.: Health Informatics Point-of-Care Medical Device Communication Part 20101: Application Profile-Base Standard (2004)

41. ISO/IEEE 11073-30300:2004.: Health Informatics Point-Of-Care Medical Device Communication Part 30300: Transport Profile-Infrared Wireless (2004)
42. ISO/IEEE 11073-20601:2010.: Health Informatics Personal Health Device Communication Part 20601: Application Profile Optimized Exchange Protocol. (2010)
43. Istepanian, R., Laxminarayan, S., Pattichis, C.S.: M-Health. Springer, New York (2006)
44. Jacques, L.B.: Electronic health records and respect for patient privacy: a prescription for compatibility. *Vand. J. Entertain. Technol. Law* **13**, 441 (2010)
45. Jha, A.K., et al.: Use of electronic health records in US hospitals. *N. Engl. J. Med.* **360**(16), 1628–1638 (2009)
46. Källander, K., et al.: Mobile health (mHealth) approaches and lessons for increased performance and retention of community health workers in low-and middle-income countries: a review. *J. Med. Internet Res.* **15**(1), e17 (2013)
47. Kotz, D.: A threat taxonomy for mHealth privacy. In: COMSNETS (2011)
48. Kulynych, J.: Legal and ethical issues in neuroimaging research: human subjects protection, medical privacy, and the public communication of research results. *Brain Cogn.* **50**(3), 345–357 (2002)
49. Li, M., Poovendran, R., Narayanan, S.: Protecting patient privacy against unauthorized release of medical images in a group communication environment. *Comput. Med. Imaging Graph.* **29**(5), 367–383 (2005)
50. Li, M., et al.: Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption. *IEEE Trans. Parallel Distrib. Syst.* **24**(1), 131–143 (2013)
51. Lin, Z., Owen, A.B., Altman, R.B.: Genomic research and human subject privacy. *Science - New York Then Washington* **305**, 183 (2004)
52. Lowrance, W.W., Collins, F.: Identifiability in genomic research. *Science* **317**, 600–602 (2007)
53. Markle Foundation.: Common framework for networked personal health information: overview and principles. Connecting For Health (2008)
54. Newcombe, H.B., et al.: Automatic linkage of vital records computers can be used to extract “follow-up” statistics of families from files of routine records. *Science* **130**(3381), 954–959 (1959)
55. Office for Civil Rights.: The HIPAA Privacy Rule and Electronic Health Information Exchange in a Networked Environment: Collection, Use, and Disclosure Limitation (2013)
56. Office for Civil Rights.: Guide to Privacy and Security of Electronic Health Information, Department of Health and Human Services (2014)
57. Office for Civil Rights.: HIPAA Privacy Rule and Sharing Information Related to Mental Health (2014)
58. Office of the National Coordinator.: Nationwide Privacy and Security Framework for Electronic Exchange of Individually Identifiable Health Information (2008)
59. Office of the National Coordinator.: Connecting Health and Care for the Nation; A Shared Nationwide Interoperability Roadmap (2014)
60. Office of the National Coordinator.: Privacy & Security Tiger Team. <http://www.healthit.gov/facas/health-it-policy-committee/hipc-workgroups/privacy-security-tiger-team> (2015)
61. Office of the National Coordinator.: Patient consent for electronic health information exchange. <http://www.healthit.gov/providers-professionals/patient-consent-electronic-health-information-exchange> (2015)
62. Office of the National Coordinator.: First annual summary of privacy and security tiger team activities: July 1, 2010 through September 30, 2013. http://www.healthit.gov/sites/default/files/privacysecurityteamannualsummarybriefing2010_2013.pdf (2015)
63. Office of the National Coordinator for Health Information Technology (ONC).: Governance Framework for Trusted Electronic Health Information Exchange (2013)
64. Office of the National Coordinator for Health Information Technology (ONC).: Federal Health Information Technology Strategic Plan, Department of Health & Human Services (2014)
65. Pajic, M., et al.: Model-driven safety analysis of closed-loop medical systems. *IEEE Trans. Ind. Inf.* **10**(1), 3–16 (2014)

66. Paszko, C., Turner, E.: Laboratory Information Management Systems. CRC Press, Boca Raton (2001)
67. Reichertz, P.L.: Hospital information systems – past, present, future. *Int. J. Med. Inform.* **75**(3), 282–299 (2006)
68. Rudin, R.S., et al.: Understanding the decisions and values of stakeholders in health information exchanges: experiences from Massachusetts. *Am. J. Public Health* **99**(5), 950 (2009)
69. Russello, G., Changyu, D., Dul, N.: Consent-based workflows for healthcare management. In: Policies for Distributed Systems and Networks, 2008. IEEE Workshop on POLICY 2008 (2008)
70. Sankararaman, S., et al.: Genomic privacy and limits of individual detection in a pool. *Nat. Genet.* **41**(9), 965–967 (2009)
71. Schimke, N., Kuehler, M., Hale, J.: Preserving privacy in structural neuroimages. In: Data and Applications Security and Privacy, vol. XXV, pp. 301–308. Springer, Berlin/Heidelberg (2011)
72. Schnell, R., Bachteler, T., Reiher, J.: Privacy-preserving record linkage using Bloom filters. *BMC Med. Inform. Decis. Mak.* **9**(1), 41 (2009)
73. Schneier, B.: Applied Cryptography: Protocols, Algorithms, and Source Code in C. Wiley, New York (2007)
74. Scholl, M., et al.: NIST SP 800 - 66 Rev1: An Introductory Resource Guide for Implementing the Health Insurance Portability and Accountability Act (HIPAA) Security Rule, National Institute of Standards and Technology (2008)
75. Shoniregun, C.A., Dube, K., Mtenzi, F.: Secure e-healthcare information systems. In: Electronic Healthcare Information Security, pp. 101–121. Springer, Berlin (2010)
76. Solo, D., Housley, R., Ford, W.: Internet X. 509 public key infrastructure certificate and CRL profile. Internet Engineering Task Force, Networking Group (1999)
77. Standards for Privacy of Individually Identifiable Health Information (PIHI), Federal Register. (codified at 45 CFR. 164.502(b)(1)) (2002)
78. Standards for Privacy of Individually Identifiable Health Information (PIHI), Federal Register. (codified at 45 CFR. 164.502(b)(2)) (2002)
79. Standards for Privacy of Individually Identifiable Health Information (PIHI), Federal Register. (codified at 45 CFR. 164.514(d)) (2002)
80. Substance Abuse and Confidentiality, Federal Register. (codified at 42 CFR. Part 2) (2014)
81. Substance Abuse and Mental Health Services Administration: Consent2Share Project. <http://www.wiki.siframework.org/SAMHSA+Consent2Share+Project> (2015)
82. Van der Linden, H., et al.: Inter-organizational future proof EHR systems: a review of the security and privacy related issues. *Int. J. Med. Inform.* **78**(3), 141–160 (2009)
83. Vest, J.R., Gamm, L.D.: Health information exchange: persistent challenges and new strategies. *J. Am. Med. Inform. Assoc.* **17**(3), 288–294 (2010)
84. Wang, Q., Hongxia, J.: An analytical solution for consent management in patient privacy preservation. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. ACM (2012)
85. West, D.: How mobile devices are transforming healthcare. *Issues Technol. Innov.* **18**(1), 1–14 (2012)
86. White, P.: Privacy and security issues in teleradiology. In: Seminars in Ultrasound, CT and MRI, vol. 25(5) (2004)
87. Wilcox, A., et al.: Architectural strategies and issues with health information exchange. In: AMIA Annual Symposium Proceedings, vol. 2006. American Medical Informatics Association (2006)
88. Zafar, A., Dixon, B.E.: Pulling back the covers: technical lessons of a real-world health information exchange vol. 129 (Pt 1), 488–492 (2007)
89. Zhang, W., et al.: Role prediction using electronic medical record system audits. In: AMIA Annual Symposium Proceedings, pp. 858–867 (2011)

Chapter 13

Managing Access Control in Collaborative Processes for Healthcare Applications

Xuan Hung Le and Dongwen Wang

Abstract Team-based patient care, biomedical research, and clinical education require coordinated access of relevant information in specific contexts of workflow and collaboration. Research on methodology development to manage information access in collaborative processes therefore is essential to build successful healthcare applications. In this chapter, we first survey the existing research on access control to support team collaboration and workflow management. We then introduce an illustrative example, New York State HIV Clinical Education Initiative (CEI), as a domain application requiring complex information access in the combined contexts of workflow and team collaboration. To address the specific challenges in access control for CEI, we present a series of studies on model development, system implementation, and effectiveness evaluation. Specifically, we describe the enhancement of the Role-Based Access Control (RBAC) model through formulating universal constraints, defining bridging entities and contributing attributes, extending access permissions to include workflow contexts, synthesizing a role-based access delegation model to target on specific objects, and developing domain ontologies as instantiations of the general model to particular applications. We illustrate the development of a generic system framework to implement the enhanced RBAC model, with three functional layers: encoding of access control policies, interpretation of these policies, and application of the policies to specific scenarios for information access management. We present an evaluation study to assess the effectiveness of the enhanced RBAC model when applied to CEI, with quantitative measures on degree of agreement with a control system as well as sensitivity, specificity, and accuracy based on a gold-standard. We close this chapter with discussions, future works, and some conclusion remarks.

X.H. Le • D. Wang (✉)

Biomedical Informatics Research & Development Center,
University of Rochester Medical Center, Rochester, NY 14642, USA
e-mail: hungfitvn@gmail.com; dongwen_wang@urmc.rochester.edu

13.1 Introduction

Patient care, medical research, and clinical education are increasingly depending on coordination and collaboration of partners from multiple disciplines in specific workflow contexts [20, 25, 26, 41, 53, 55, 57, 64, 69, 72, 83]. Information systems can be effectively used to facilitate team collaboration and workflow management [5, 8, 10, 11, 15, 17, 31, 32, 38, 43, 45, 56, 63, 73, 76, 88]. For an information system to present the right information to the right people at the right time, access control is a critical requirement.

Traditionally, access control is used to limit the access to certain information to protect patient privacy [6, 14, 58, 74], to ensure confidentiality of sensitive data [2, 46, 77], and to filter out irrelevant information in order to reduce information overload [3, 33, 54]. In collaborative processes, access control can also facilitate information sharing to address information needs [15, 25, 55], to improve team collaboration [8, 10, 26, 38, 43, 53, 69, 88], and to enhance workflow management [5, 17, 63, 83]. Previous research has shown that the requirements of information access are continuously changing. Such changes, either in scope of information (i.e., availability of data vs. restriction of access to certain records) or level of access (i.e., review vs. documentation), depend on the specific context of workflow (i.e., goals, tasks, and available resources) and particular requirements of team collaboration (i.e., expertise needed for work and responsibilities assigned to individual team members) [25, 30, 55, 69, 70, 83, 88]. We also note that the requirements for workflow management and team collaboration frequently interfere with each other. For example, the tasks in a specific workflow context may need particular expertise from team members. Only considering one aspect of workflow or team collaboration in access control will not be able to handle the dynamic nature of team collaboration and the complexity of workflow involving multiple team members. Addressing the needs of information access management in the combined context of team collaboration and workflow is thus becoming a fundamental requirement to all information systems.

In this chapter, we first review the existing research on access control in team collaboration and workflow management. We then introduce the New York State HIV Clinical Education Initiative (CEI) [61] as an illustrative example. We describe our approaches to enhance the Role-Based Access Control (RBAC) model [51], followed by a generic system framework for its implementation [49]. We present an evaluation study to assess the effectiveness of the enhanced RBAC model when applied to CEI [52]. We close the chapter with discussions of the results, future works, and conclusion remarks.

13.2 Related Works

Since the late 1960s, the research community has proposed various information access control models. Among these, Access Matrix [47] was the first introduced by

Lampson, with a simple structure consisting two elements: *access permission* and *user identification*. In 1973, Lapadule and Bell proposed Rule-Based Access Control [48], within which access permissions were assigned to users based on specific rules. Later, the United States Department of Defense released Discretionary Access Control (DAC) [16] and Mandatory Access Control (MAC) [16]. In these models, access permissions were managed based on individual users, which introduced complexity and cost to manage large-scale systems. Attribute-Based Access Control (ABAC) [89] is based on logical combination of the attributes held by subjects and objects, as well as the environment conditions. Since the attributes can be flexibly defined based on any parameters, ABAC is extremely powerful in its expressiveness. RBAC [19] was proposed in the early 1990s and then standardized by the National Institute of Standards and Technology (NIST). RBAC solved the complexity issue by assigning permissions to users based on their roles in organizations, to denote specific job functions and the associated authorities and responsibilities. For two decades, RBAC has been widely used owing to its salient features such as generalization, simplicity, and effectiveness.

With regard to implementation, these models and their extensions have been applied to specific applications to manage information access. System frameworks have been developed to implement the core components of access control models that are independent of specific domains or applications [1, 12, 27, 40, 62]. Nevertheless, none of these models or approaches is sufficient for managing information access in the combined context of team collaboration and workflow.

Because of special domain needs, healthcare has unique requirements on information access management. A number of access control applications have been developed to protect patient privacy [7, 28, 68], to facilitate access to relevant information by healthcare providers [44, 50, 75], and to ensure appropriate level of access to electronic medical records and personal health records [13, 60, 79]. Regarding the specific application contexts, Predeschly et al. introduced important concepts to ensure information security in adaptive processes [70]; however, they did not address the issue of team collaboration. Grando et al. addressed both process management and team work [30], but only from the prospective of workflow tasks.

Research on using information systems to facilitate communication in a collaborative environment to accomplish complex tasks, the so-called computer-supported cooperative work (CSCW), is very active [8, 38, 43, 69, 88]. Previous efforts to apply CSCW to healthcare applications have focused on support of collaborative clinical tasks [30, 43, 69], facilitating communication among team members of biomedical research [8, 26, 53], and coordination of healthcare management activities [10, 38, 88]. Managing information access is a fundamental requirement in collaborative environments and has resulted to the development of access control models in such contexts [29, 39].

With regard to facilitating workflow, it is widely documented in the literature that this is one of the most important factors for a successful implementation of information systems for biomedical research, clinical education, and patient care [5, 17, 63, 83]. Since the workflow context defines the specific tasks that

require access to particular information, managing information access in context of workflow is an essential requirement to information systems [70].

Among the various models and frameworks, ABAC and RBAC worth special mentioning. With its expressiveness, ABAC has the potential to be used for information access management in a collaborative workflow. However, without a structure to address specific workflow contexts and team collaboration requirements, it can easily become too complex to manage. RBAC can significantly improve simplicity of management by introducing roles, which could be very helpful for team collaboration. Meanwhile, it can still address complex requirements through the constraints. In a sense, RBAC can be considered as a specialization of ABAC—it emphasizes the attribute of role so much that it singles it out as a predefined component. Our model development, for example, formulation of special types of constraints, definition of bridging entities and contributing attributes, and extension of access permission to include workflow status, is based on this philosophy. In particular, these unique features required for workflow management and team collaboration are specified as predefined components such that they can be used to facilitate information access management in collaborative processes (see Sect. 13.4 below).

13.3 An Illustrative Example: New York State HIV Clinical Education Initiative

Our general approach to managing information access in collaborative processes includes three components, i.e., *model development*, *system implementation*, and *effectiveness evaluation*. Before presenting the technical details, we first introduce the New York State HIV CEI [61] as an illustrative example. This example will be used as a specific application to explain the concepts when describing our approaches in the following sections.

The CEI program is sponsored by the New York State Department of Health AIDS Institute. It has been engaging in clinical education for two decades to meet the needs of community clinicians who provide healthcare to HIV patients in New York State. In 2008, the CEI program was re-organized to reflect the new priorities in specific topics of HIV clinical education, to address the differences between the Upstate New York region and the Metropolitan New York City area, to develop online training programs, and to provide resource coordination and program evaluation. In consequence, the CEI program created seven training centers:

1. Testing, Post-Exposure Prophylaxis, and Diagnosis Center (TPDC)
2. Prevention and Substance Use Center (PSUC)
3. Mental Health Center (MHC)
4. Clinical Education Center for Upstate Providers (CECUP)
5. Technology Center (TC)
6. Resource and Referral Center (RRC)
7. Evaluation Center (EC)

Table 13.1 The responsibilities of the CEI centers^a

Center	Role	Responsibility	Topic	Location	Format
CECUP	r_{cecup}	Onsite training	All topics	Upstate region	Onsite
TPDC	r_{tpdc}	Onsite training	Testing, PEP, diagnosis	All areas	Onsite
PSUC	r_{psuc}	Onsite training	Prevention, substance use	All areas	Onsite
MHC	r_{mhc}	Onsite training	Mental health	All areas	Onsite
TC	r_{tc}	Online training	All topics	All areas	Online
RRC	r_{rrc}	Resource and referral	All topics	All areas	Onsite/online
EC	r_{ec}	Evaluation	All topics	All areas	Onsite/online

^aTable reprinted from [51], with permission from Elsevier

Here each CEI Center is in charge of a range of training activities, while in everyday operation a specific training session may require expertise and resources from multiple CEI Centers. Specifically, four training centers, i.e., CECUP, TPDC, PSUC, and MHC, are charged to provide onsite trainings. The responsibilities of each center are defined by specific training topics and geographical locations. In particular, the three specialty training centers (TPDC, PSUC, and MHC) are charged to provide training on specific topics of HIV clinical education for the entire state. For example, TPDC's expertise includes HIV testing, post-exposure prophylaxis, diagnosis, and acute HIV infection; PSUC's expertise includes HIV prevention and substance use; MHC's expertise includes all mental health related issues. In contrast, CECUP is in charge of all training topics (including the specialty topics described above and general topics such as adult care, pediatric care, oral health, pharmacy, ethical/legal issues, and clinical trials) in upstate New York. Within the CECUP Center, three catchment areas, i.e., Albany Medical Center (AMC), University of Rochester Medical Center (URMC), and Erie County Medical Center (ECMC), are further segmented. In addition to onsite training, TC is in charge of online training; RRC is in charge of program resources and referrals; EC is in charge of program evaluation. Table 13.1 is a summary of the responsibilities of each CEI Center.

When a healthcare organization in New York State requests a training session, it can select from a list of courses or topics, each of which is mapped to specific CEI Centers. Depending on the location of the requester (where the onsite training will be delivered), the selected courses or topics, and the preferred training format, the CEI Centers will work individually or collaboratively to deliver the training service.

For a specific training session in CEI, the typical training workflow consists of several major stages: training requested by a healthcare organization (*request_received*), calling back by a CEI staff and training arrangement pending (*arrangement_pending*), scheduling of training event (*training_scheduled*), and completion of training (*training_completed*). If a training session is progressing as planned, it will pass through these stages step-by-step. From time to time, a training session may step back to an earlier stage (for example, when a scheduled training session is cancelled); a process may also terminate earlier before the training is actually delivered (for example, when a request is beyond the scope of CEI and referred to other training programs).

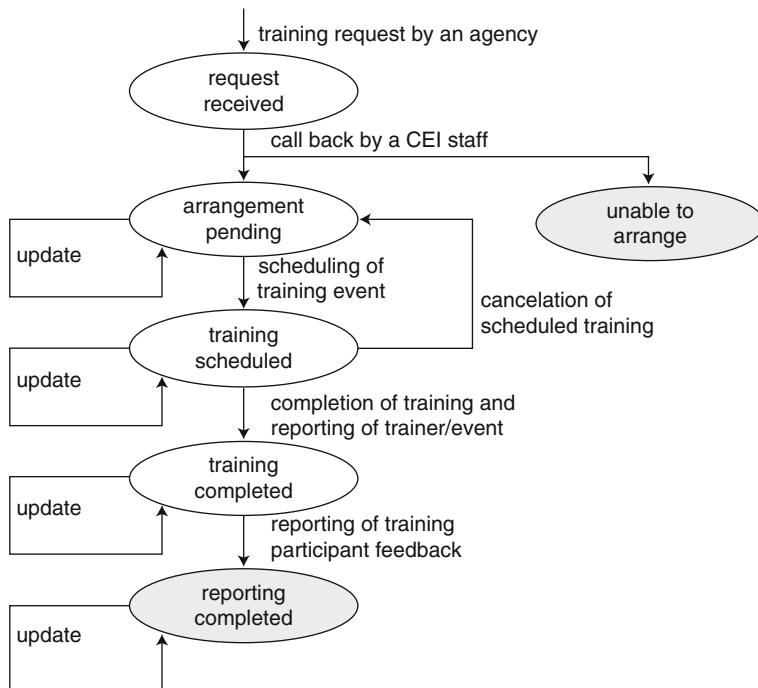


Fig. 13.1 Workflow of a CEI training session (reprinted from [51], with permission from Elsevier)

The stages of a training session can be represented in a state-transition diagram, as shown in Fig. 13.1. In each stage of the process, a CEI Center staff can review and document specific information for the training session that she/he is charge of. In case a training session involves collaboration of multiple CEI Centers, this information is also available for review by staff from the collaborating centers. To manage the CEI training program, we have developed the CEIAdmin system for CEI Centers to review the relevant training requests, to document training activities over the training process, and to collaborate with other CEI Centers. Additional details of the CEIAdmin system can be found in [51].

13.4 Development of the Enhanced RBAC Model

To develop an access control model to define access permissions in the context of team collaboration and workflow management, we established the following general principles: (1) *effectiveness*: the model could be applied to define the exact scope and level of information access in collaborative processes; (2) *simplicity*: the model should have a relatively simple structure and leverage common representation formats so it could be easily implemented; and (3) *generalizability*: the model

should initially serve the purpose of information access management for CEI and eventually be able to generalize to other domains and applications. With these general principles in mind, we reviewed the existing models for access control and workflow management. Among the reviewed models, we found that RBAC [19] is widely used, has a simple structure, and its internal constructs can be extended for new requirements. For these reasons, we selected to start with it to develop the proposed model as an enhancement of RBAC.

13.4.1 Overview of the Enhanced RBAC Model

The enhanced RBAC model extends the core RBAC through: (1) formulation of specific types of *universal constraints* to bind on user-role assignments, role-permission assignment, and access permissions; (2) definition of *bridging entities* and *contributing attributes* to support access management in collaborative environment; (3) extension of access permissions to include *workflow context*; and (4) synthesis of a *role-based access delegation* model targeting on specific objects to balance between flexibility and need-based access.

The universal constraints are collections of constraints that regulate specific aspects of access control. As shown in Fig. 13.2, these constraints could be *separation of duty* (as defined in the core RBAC [19]), *access delegation* (to delegate access permission under specific conditions), *collaboration constraint* (to support team collaboration), *temporal constraint* (to facilitate workflow management), and *organizational constraint* (to define roles based on organizational structure). In association with the *universal constraints*, we propose an *enhanced permission set*. Instead of using a simple structure of *asset-action* pair, as defined in the core RBAC model, we formulate the permission set within the context of a specific *workflow status*. Here a workflow status represents specific tasks or goals defined as part of a workflow, the execution of which indicates the current state of the process. The *workflow status*, together with *action* and *asset* defined in the core RBAC, are then bundled with particular types of *universal constraints* to define access permissions for team collaboration and workflow management.

To implement the enhanced RBAC model in a specific domain or application, we introduce the concept of *domain ontologies*. The domain ontologies define: (1) the specific instances and specializations of the general model components (i.e., users, roles, objects, operations, workflow statuses, and universal constraints) for implementation in a particular domain or application; (2) the relations among domain components; and (3) the constraints bound on domain components and relations. For example, when defining the domain ontologies for the CEI project, we have specified three types of collaboration constraints, i.e., *geographical constraint* (to manage collaboration based on geographical areas), *topic constraint* (to manage collaboration based on training topics), and *format constraint* (to manage collaboration based on training format). It is important to note that the domain ontologies are

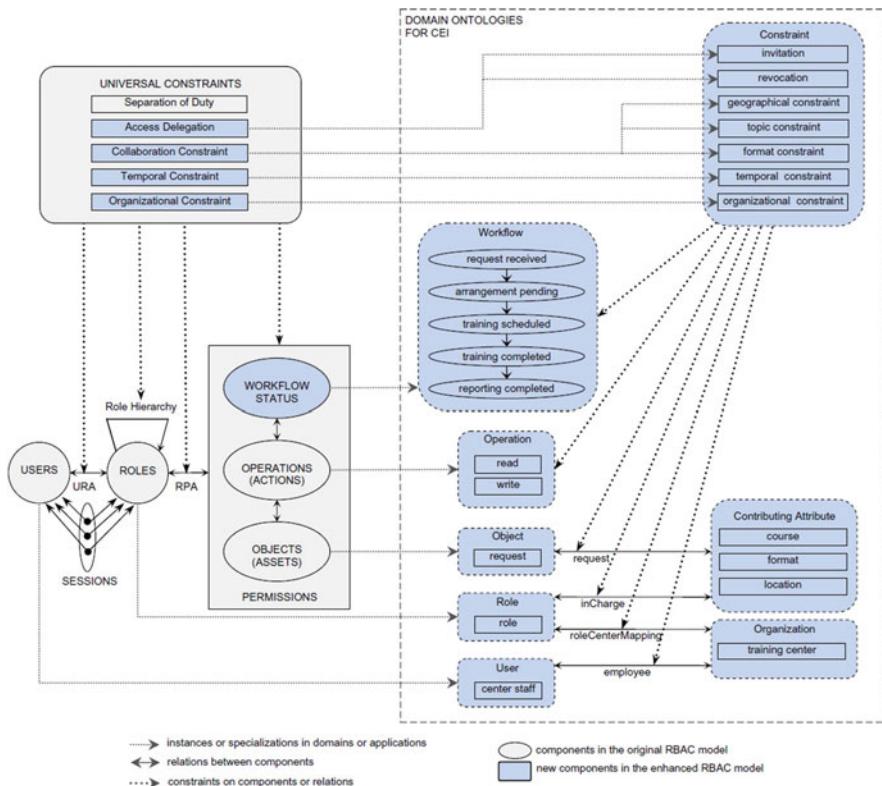


Fig. 13.2 Enhanced RBAC model with universal constraints, workflow in permissions, and domain ontologies (reprinted from [51], with permission from Elsevier)

open, configurable structures. Depending on specific applications, various types of components, relations, and constraints can be included to meet domain requirements without breaking the core RBAC structure.

13.4.2 Support Team Collaboration: Bridging Entities and Contributing Attributes

To support information access management in the context of team collaboration, an essential requirement is to specify the constraints on access permissions based on the attributes that define the collaborative responsibilities. These attributes are used to divide the collaborative tasks to individual team members and to re-assemble them together once these tasks are completed. For coordination of clinical education, a collaborative work may involve learners (medical students, nursing students, etc.) and teachers (faculty members from various healthcare related disciplines).

A collaboration may simply happen between the two parties of a learner and a teacher, who are associated to each other through a training session. We name this entity that connects the collaborating parties (for example, a training session) a *bridging entity*.

A collaboration may also involve multiple learners and teachers, each of whom is engaging in specific activities of a training session. For example, in a training session of the CEI program, each collaborating CEI Center contributes its expertise and resources, as defined by specific training topics, formats, and locations; by integrating the expertise and resources from all collaborating CEI Centers, we can accommodate the various needs by a specific training session. We name these contributions that partition the bridging entity (for example, training topics, formats, and locations) *contributing attributes*-eventually, when each party finishes its own activities, defined by these attributes, it contributes to the overall collaborative work.

When implementing these concepts, the definition of contributing attributes can be based on multiple dimensions, each of which represents a specific type of universal constraint. The contributions from an individual collaborating party can then be specified as a logical combination of multiple constraints. With these specifications, the collaborating parties will satisfy the defined constraints and thus have access to the relevant training information, while the other parties not directly involved in the collaboration will not satisfy these constraints and thus have no access permissions. The relationships among the bridging entity, contributing attributes, and collaborating parties are shown in Fig. 13.3.

With these definitions, we can formally represent topic constraint, format constraint, and geographical constraint through the following predicates:

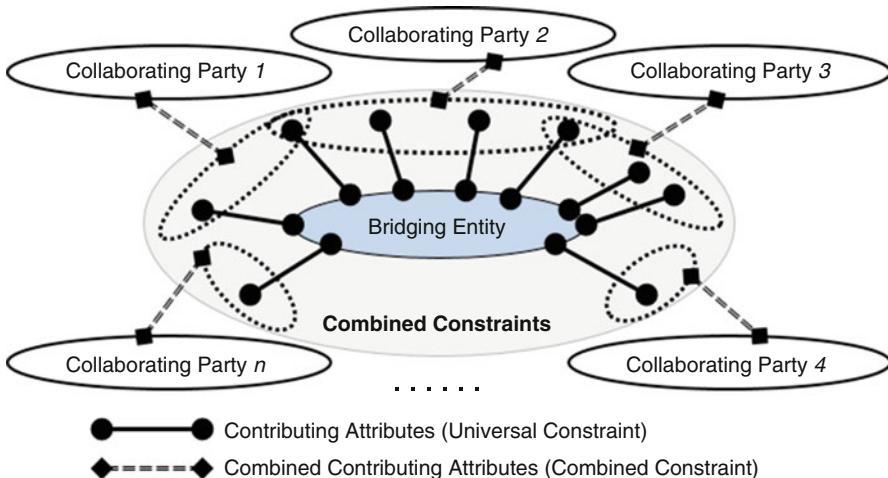


Fig. 13.3 Collaboration model with bridging entity and contributing attributes (reprinted from [51], with permission from Elsevier)

<code>inChargeCourse(r, course)</code> :	role r is in charge of course $course$.
<code>inChargeFormat(r, format)</code> :	role r is in charge of format $format$.
<code>inChargeLocation(r, loc)</code> :	role r is in charge of location loc .

A complete representation model in first-order predicate logic can be found in [51]. In Sect. 13.4.6, we will use specific examples from the CEI project to illustrate how to use this model to address specific requirements for information access management in a collaborative environment.

13.4.3 Extending Access Permissions to Include Workflow Contexts

As briefly discussed in Sect. 13.4.1, the core RBAC model defines access permissions as *asset-action pairs*. The proposed enhancement of RBAC introduces *workflow status* in representation of permissions such that we can define access authority within specific context of workflow. To simplify the model, we select to not include any technical details to handle workflow management. Instead, we assume that a separate workflow engine outside of the access management model will be used for this purpose [81, 87]. Inside the access control model, we only include a specific workflow status, within which the access permissions are specified. Formally, we define the following problem domains:

OBJECT:	a set of system assets or resources.
OPERATION:	a set of actions or operations.
WORKFLOWSTATUS:	a set of workflow statuses.

We can then define access permissions in context of workflow management as:

$$\text{PERMISSION} = \text{OBJECT} \times (\text{WORKFLOWSTATUS} \times \text{OPERATION})$$

In other words, if $p \in \text{PERMISSION}$, $\text{obj} \in \text{OBJECT}$, $w \in \text{WORKFLOWSTATUS}$, and $\text{op} \in \text{OPERATION}$, where $p = (\text{obj}, w, \text{op})$, it means operation op on object obj within the context of workflow status w is permitted. The relationship among access permission, object, workflow status, and operation is shown in Fig. 13.2. The complete model defined in first-order predicate logic can be found in [51].

13.4.4 Role-Based Access Delegation Targeting on Specific Objects: Providing Flexibility for Access Control in Collaborative Processes

Access delegation regulates the situation when an authorized user transfers parts or all of his/her access permissions to another user who is otherwise not authorized

to have such access. It is an indispensable component for information access management in a collaborative environment, as it can enable access under certain conditions to facilitate team work.

The proposed access control model supports access delegation through explicitly inviting other parties (*invitee*) for collaboration and granting them the appropriate level of access permissions that the inviting party (*inviter*) holds. Traditionally, access delegation under RBAC concerns about the process that the inviter grants his/her access permission to the invitee through delegating the former's role to the latter [90, 91]. These approaches only differentiate the original roles from the delegated roles, but treat all objects in the permissions uniformly [91]. In the enhanced RBAC model, since our primary purpose to delegate access permission is to enable collaboration in specific workflow context, the delegation has to target on specific objects in that context. For example, when managing a collaborative training program, a training center can invite another center to collaborate and such collaboration only happens when they work together on specific training sessions. Thus, delegation in this context needs to be defined based on specific objects, i.e., the training sessions under the collaboration. To address this requirement, the proposed delegation model is focusing on the situation that a role (represented by a specific user) delegates certain access permission on specific objects or resources to another role. In other words, this is a *role-to-role delegation* that targets on only permissions defined for specific objects. The reverse concept of access delegation is *access revocation*, which deprives a party of specific access permissions that were granted previously through access delegation. Similar to access delegation, access revocation in the proposed model is defined between two roles and targeted on access permissions for specific objects. To formally represent access delegation and revocation, we use the following predicates:

role r_{inviter} invites role r_{invitee}
to collaborate on object obj and

invitation($\text{obj}, \text{op}, r_{\text{invitee}}, r_{\text{inviter}}$): grants r_{invitee} the permission to
perform operation op in this
context.

role r_{revoker} deprives role
 r_{revoked} of the permission to
perform operation op with regard
to object obj .

revocation($\text{obj}, \text{op}, r_{\text{revoked}}, r_{\text{revoker}}$): grants r_{revoked} the permission to

It is important to note that only the delegated permissions can be revoked. Certain access permissions are defined originally by the contributing attributes (see Sect. 13.4.2), and therefore those permissions held by the related collaborating parties cannot be removed through a revocation. We name these collaborating parties that hold access permissions through the original assignment of authorizations *root parties* (essentially equivalent to the original users in a conventional RBAC delegation model when team collaboration is not a concern [90]). In a collaborative environment, a specific party may play a leading role to orchestrate the collaboration among all parties. We name this specific party the *leading party*.

For example, in the CEI project the leading party (a CEI Center) of a specific training session needs to have the authority to invite other parties (CEI Centers) for collaboration, to document training related issues, and to update the recorded data. Therefore it holds the “*write*” permission (document, update, invite other centers, etc.) for the data related to that specific training session.

To model these features, the CEI policy for access delegation requires that:

1. There is only one leading party at any time.
2. The leading role is transferable, which means a leading party can transfer/delegate its leading role to another party (and immediately becomes a non-leading party due to the requirement defined above).
3. Only the leading party can delegate (invite) and revoke access permissions.
4. Certain access permissions held by the root parties through definitions of contributing attributes cannot be revoked.

A complete representation of access delegation and revocation in first-order predicate logic can be found in [51].

13.4.5 Integration of Multiple Representation Elements for Definition of Universal Constraints

As described above, information access in the specific context of team collaboration and workflow management is typically around the bridging entity and the contributing attributes. Depending on specific applications, the representation of bridging entity, contributing attributes, and access permissions may need a complex structure. For example, in the CEI project the bridging entity is a specific training session, which is triggered by a training request from a healthcare organization located in New York State. A training request contains the specific information about the requested course or training topics, as well as the format of training. The location of the training can be derived from the requesting organization, where the training session is typically delivered. Depending on the training topic, format, and location, specific CEI Centers with the resources and expertise will work collaboratively on a training session. As the process moves forward, the workflow status of the training session is changing. Therefore, the access permissions, which are specified based on the workflow status, are also redefined. To integrate all these requirements, we define the following predicates:

request(req, course, format, agency):	agency agency makes a request req for a course course in the format format.
---------------------------------------	-----------------------------------------------------------------------------------

location(agency,loc):	agency agency is located at the location loc.
status(req,w):	request req is processed and its current status in the workflow is w.

Using a logical combination of these predicates together with those defined previously in Sects. 13.4.2 and 13.4.4, we can complete the constraint definition. For example, the constraints on assignment of “read” permission (review, query, etc.) to roles for the CEI program collaboration can be defined as:

```

 $\forall r \in \text{ROLE}, \forall \text{req} \in \text{REQUEST}, \forall w \in \text{WORKFLOWSTATUS},$ 
 $\forall \text{opr}_r \in \text{OPERATION}_{\text{READ}}, \forall \text{op}_w \in \text{OPERATION}_{\text{WRITE}}, \forall \text{course} \in$ 
 $\text{COURSE}, \forall \text{format} \in \text{FORMAT}, \forall \text{agency} \in \text{AGENCY}, \forall \text{loc} \in$ 
 $\text{LOCATION},$ 
 $\exists r_{\text{inviter}} \in \text{ROLE}(r_{\text{inviter}} \neq r) :$ 
 $\text{role-permission}(r, \text{req}, w, \text{opr}_r) \leftarrow$ 
 $\text{status}(\text{req}, w) \wedge$ 
 $\{ [\text{request}(\text{req}, \text{course}, \text{format}, \text{agency}) \wedge$ 
 $\text{location}(\text{agency}, \text{loc}) \wedge$ 
 $\text{inChargeLocation}(r, \text{loc}) \wedge$ 
 $\text{inChargeCourse}(r, \text{course}) \wedge$ 
 $\text{inChargeFormat}(r, \text{format})] \vee$ 
 $[\text{invitation}(\text{req}, \text{opr}_r, r, r_{\text{inviter}}) \wedge$ 
 $\text{role-permission}(r_{\text{inviter}}, \text{req}, w, \text{op}_w)] \}$ 

```

These constraints essentially define the following rules for information access management:

1. The “read” permission depends on the requested course, training format, location of the requesting organization, and workflow status of the training session.
2. The “read” permission is granted to only the roles (CEI Centers) that are in charge of the training location, course, and format (assignment of access permission for the root parties).
3. The “read” permission can be granted to a role (CEI Center) through delegation (invitation) from another role (CEI Center) that holds the authority (the leading center).

It is important to note that these predicates and constraints are defined for the specific requirements of the CEI project. When applying the enhanced RBAC model to another domain or application, we can introduce a different set of domain ontologies, but still keep the general framework of the model intact.

13.4.6 Case Studies to Encode Access Policies for CEI

13.4.6.1 User, Roles, Objects, and Access Permissions

As shown in Table 13.1, the definitions of training expertise and responsibilities are based on individual CEI Centers. Staff from a specific CEI Center have the same level of access permissions to the training data. Therefore, defining each CEI Center as a unique role is a natural selection. For example, we can define a unique role “ r_{cecup} ” for CECUP; all staff working for CECUP are then automatically assigned to this role. In a formal representation with the enhanced RBAC model, this can be expressed as a constraint on the user role assignment:

```
 $\forall u \in \text{USER}, \forall r \in \text{ROLE}, \forall \text{org} \in \text{ORGANIZATION}:$ 
\text{user-role}(u, r) \leftarrow
\text{employee}(u, \text{org}) \wedge \text{roleCenterMapping}(r, \text{org})
```

Thus, if staff “A” works for “CECUP”, she is automatically assigned to the “ r_{cecup} ” role:

```
“A”  $\in \text{USER}$ , “ $r_{cecup}$ ”  $\in \text{ROLE}$ , “CECUP”  $\in \text{ORGANIZATION}$ :
\text{user-role}("A", "r_{cecup}") \leftarrow
\text{employee}("A", "CECUP") \wedge \text{roleCenterMapping}("r_{cecup}", "CECUP")
```

When a specific training request is entered into the CEIAdmin system, it becomes a system resource (object), upon which access permissions can be defined. These permissions regulate whether a user in a particular role is allowed to perform specific operations on these system resources. For example, “staff from TPDC (in role “ r_{tpdc} ”) have “read” permission to training request “1090” (which is in workflow status “request-received”)” can be formally represented as:

```
“ $r_{tpdc}$ ”  $\in \text{ROLE}$ , “1090”  $\in \text{REQUEST}$ ,
“request-received”  $\in \text{WORKFLOWSTATUS}$ , “read”  $\in \text{OPERATION}_{\text{READ}}$ :
\text{role-permission}("r_{tpdc}", "1090", "request-received", "read")
```

With the role-permission assignments that we defined here, we can specify universal constraints to regulate CEI Center collaboration, training workflow, and access delegation.

13.4.6.2 Collaboration Among CEI Centers

When a training session requires expertise and resources from multiple CEI Centers, the related CEI Centers need to work together to deliver the training service collaboratively. As described in Sect. 13.4.2, we can define geographical, topic, and format constraints to bind on the access permissions to the related training requests. These constraints essentially specify the contributing attributes of the bridging entity. For example, “Eastman-Dental-Center”, a healthcare organization located in “Monroe” (a county in Upstate New York region), has requested an “onsite” training session (request “1090”, currently in workflow status “request-received”) with the course “Acute-HIV-Infection”. In this case, a collaboration will become necessary between CECUP and TPDC, as both centers can satisfy the constraints (see responsibilities defined in Table 13.1). Thus, if Staff “A” is working for CECUP and Staff “P” is working for TPDC, both “A” and “P” have “read” access permission to this training session. These requirements can be formally represented as:

```

“A” ∈ USER, “rcecup” ∈ ROLE, “CECUP” ∈ ORGANIZATION:
user-role(“A”, “rcecup”) ←
    employee(“A”, “CECUP”) ∧ roleCenterMapping(“rcecup”, “CECUP”)

“rcecup” ∈ ROLE, “1090” ∈ REQUEST,
“request-received” ∈ WORKFLOWSTATUS, “read” ∈ OPERATIONREAD,
“Acute-HIV-Infection” ∈ COURSE, “onsite” ∈ FORMAT,
“Eastman-Dental-Center” ∈ AGENCY, “Monroe” ∈ LOCATION:
role-permission(“rcecup”, “1090”, “request-received”, “read”) ←
    status(“1090”, “request-received”) ∧
    request(“1090”, “Acute-HIV-Infection”, “onsite”, “Eastman-Dental-Center”) ∧
    location(“Eastman-Dental-Center”, “Monroe”) ∧
    inChargeLocation(“rcecup”, “Monroe”) ∧
    inChargeCourse(“rcecup”, “Acute-HIV-Infection”) ∧
    inChargeFormat(“rcecup”, “onsite”)

“P” ∈ USER, “rtpdc” ∈ ROLE, “TPDC” ∈ ORGANIZATION:
user-role(“P”, “rtpdc”) ←
    employee(“P”, “TPDC”) ∧ roleCenterMapping(“rtpdc”, “TPDC”)

“rtpdc” ∈ ROLE, “1090” ∈ REQUEST,
“request-received” ∈ WORKFLOWSTATUS, “read” ∈ OPERATIONREAD,
“Acute-HIV-Infection” ∈ COURSE, “onsite” ∈ FORMAT,
“Eastman-Dental-Center” ∈ AGENCY, “Monroe” ∈ LOCATION:
role-permission(“rtpdc”, “1090”, “request-received”, “read”) ←

```

(continued)

```

status("1090","request-received") ∧
request("1090","Acute-HIV-Infection","onsite",
      "Eastman-Dental-Center") ∧
location("Eastman-Dental-Center","Monroe") ∧
inChargeLocation("r_tpdc","Monroe") ∧
inChargeCourse("r_tpdc","Acute-HIV-Infection") ∧
inChargeFormat("r_tpdc","onsite")

```

Meanwhile, staff from other CEI Centers would not see this request as they cannot satisfy the constraint above and therefore do not have access permissions.

13.4.6.3 Management of Training Workflow

As described earlier, access permissions to a specific training session is based on its workflow status. Thus, by verifying that request “1090” is in status “request-received”, we can define access permission for this stage of the training workflow (see the logic statements above). If the workflow is moving forward to the “arrangement-pending” stage and all other conditions remain as the same, we will be able to derive the access permission for the new workflow status through examination of the constraints:

```

“r_cecup” ∈ ROLE, “1090” ∈ REQUEST,
“arrangement-pending” ∈ WORKFLOWSTATUS,
“read” ∈ OPERATIONREAD, “Acute-HIV-Infection” ∈ COURSE,
“onsite” ∈ FORMAT, “Eastman-Dental-Center” ∈ AGENCY,
“Monroe” ∈ LOCATION:
role-permission(“r_cecup”,“1090”,“arrangement-pending”,“read”) ←
    status(“1090”,“arrangement-pending”) ∧
    request(“1090”,“Acute-HIV-Infection”,“onsite”,
            “Eastman-Dental-Center”) ∧
    location(“Eastman-Dental-Center”,“Monroe”) ∧
    inChargeLocation(“r_cecup”,“Monroe”) ∧
    inChargeCourse(“r_cecup”,“Acute-HIV-Infection”) ∧
    inChargeFormat(“r_cecup”,“onsite”)

```

13.4.6.4 Inviting other CEI Centers for Collaboration

Under certain scenarios, we need the flexibility to invite a specific CEI Center to participate in a training session, even though this center does not need to involve

according to the original definition of responsibilities. For example, the request “1090” is made by a healthcare organization located in Upstate New York region for an onsite training on the topic of “Acute-HIV-Infection”. According to the contributing attributes, only CECUP and TPDC are required to collaborate on this training session. When scheduling this particular training session, if it is decided that certain mental health related issues will also be addressed in addition to the main topic, MHC will need to participate in this session. In this case, the leading center (suppose it is CECUP, with “write” permission for this request) can invite MHC to collaborate on this specific training session. This scenario can be formally represented as:

```

“rmhc” ∈ ROLE, “rcecup” ∈ ROLE, “1090” ∈ REQUEST,
“training-scheduled” ∈ WORKFLOWSTATUS, “read” ∈ OPERATIONREAD
“write” ∈ OPERATIONWRITE, “Acute-HIV-Infection” ∈ COURSE,
“onsite” ∈ FORMAT, “Eastman-Dental-Center” ∈ AGENCY,
“Monroe” ∈ LOCATION:
role-permission(“rmhc”, “1090”, “training-scheduled”, “read”) ←
    status(“1090”, “training-scheduled”) ∧
    invitation(“1090”, “read”, “rmhc”, “rcecup”) ∧
    role-permission(“rcecup”, “1090”, “training-scheduled”, “write”)

```

13.5 System Framework for Implementation of Enhanced RBAC

Once the access policies are defined for specific applications based on the enhanced RBAC model, we need to implement them to ensure that these policies are correctly interpreted and applied, such as to grant particular users in certain roles the appropriate scope and level of information access to support workflow management and team collaboration. In order to generalize to multiple domains and applications, it is essential to develop a system framework for implementation of the core components of the enhanced RBAC.

In this section, we describe the development of such a system framework, which includes three functional layers:

1. Encoding of access control policies.
2. Interpretation of the encoded policies.
3. Application of the policies to specific cases and scenarios for information access management [49].

With these functional layers, the system framework can provide a flexible platform to develop and to implement policies for information access management in col-

laborative processes. To support the management of information access, to simulate the application of access control policies, and to present the access permissions in specific cases and scenarios, we have developed a demonstration tool with two primary functions: (1) selecting the combinations of users, roles, objects, operations, and workflow statuses; and (2) displaying the associated constraints and access permissions. Again, we will use CEI as a specific example to illustrate the use of this system framework for implementation of information access control policies to facilitate collaborative processes.

13.5.1 System Architecture

To implement the enhanced RBAC model, we designed a system architecture with three layers, specified as follows:

Layer 1: Policy Encoding Layer In this layer, policies regarding information access control for a specific application can be defined.

Layer 2: Policy Interpretation Layer In this layer, the encoded access policies are interpreted based on specific combinations of users, roles, objects, operations, and workflow statuses as well as the universal constraints bound on them.

Layer 3: Access Control Application Layer In this layer, the encoded access control policies are applied to specific cases or scenarios to make decisions on granting or denying access.

As illustrated in Fig. 13.4, these three layers are functioning independently but meanwhile closely integrated to each other through information flows. Specifically, Layer 1 is where the system administrators specify access control policies through a policy editor. The outputs from this process are the encoded policies that are fed to Layer 2 once they are validated. Layer 2 is where the encoded policies are interpreted in specific application contexts. During this process, it requires particular application contextual information that comes from Layer 3. The results of policy interpretation in Layer 2 are then fed to Layer 3, where decisions on granting or denying access to specific cases and scenarios are made.

To support policy encoding with the continuous development of the enhanced RBAC model, we have employed a three-level encoding schema to differentiate the core RBAC model, the model extension, and the policy instances. To interpret the encoded policies, we have leveraged a Java-based rule engine that can directly retrieve and process the encoded access policies from the policy encoding layer. To handle the workflow management, we assume there is an external workflow engine that can be integrated with the policy interpretation layer. Finally, the application layer implements system interfaces to retrieve application contextual information such as users, roles, objects, and operations. We describe the details of each layer in the following sections.

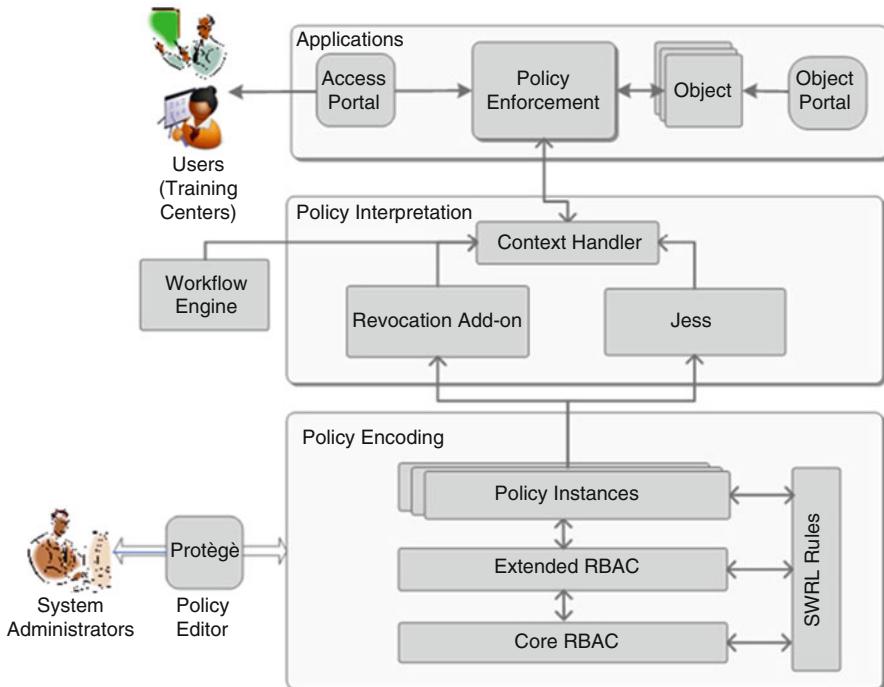


Fig. 13.4 System architecture (reprinted from [49])

13.5.2 Encoding of Access Policies

To encode access control policies, we adopted an existing tool, Protégé (version 3.4.7) [71], as the editor. Protégé provides a flexible platform and a rich technical environment for knowledge acquisition and ontology development. We selected it as the editor mainly because of three reasons:

1. It supports a layered approach for model definition, which is helpful for continuous development of the enhanced RBAC model.
2. It differentiates ontological models from instances, which is a nice feature to support the development of the general enhanced RBAC model and the implementation of the specific access control policies when applied to particular domains and problems.
3. It incorporates many add-ons, in particular, the Semantic Web Rule Language (SWRL) [66, 80] and the jess rule engine [23, 37], which can be adopted for encoding and interpretation of the universal constraints.

Encoding of specific access control policies is based on the enhanced RBAC model. To build the enhanced RBAC in Protégé, we employed a two-level structure, with the core RBAC model as the basis and the extended model encoded on top of it. The

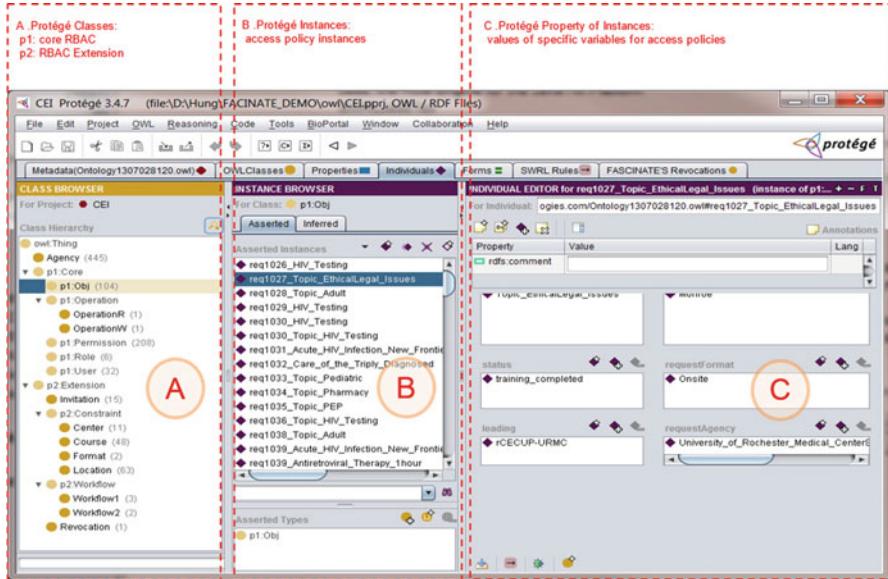


Fig. 13.5 Three-level access control policy encoding in Protégé (reprinted from [49])

core RBAC schema defines five basic components of the RBAC model, including users, roles, permissions, operations, and objects. The extended model integrates additional representation elements such as universal constraints, workflow statuses, and constructs that are used for access delegation and revocation. With this two-level structure, we can easily support the continuous development of the enhanced RBAC model, for example, through defining new types of universal constraints.

Once the enhanced RBAC is built, we can encode access control policies as instances of particular model elements. These policies can be grouped together based on individual or classes of applications. Figure 13.5 is a screenshot of using the Protégé tool to encode the enhanced RBAC model and the access control policy instances for the CEI project.

When defining access control policies for specific applications, we typically encode these policies as universal constraints. In the enhanced RBAC model, the universal constraints are defined in first-order predicate logic [51]. To encode these constraints in Protégé, we adopted a Protégé add-on that incorporates a SWRL editor and a jess rule engine for rule execution [18]. A SWRL rule contains an antecedent part, which is referred to as the body of the rule, and a consequent part, which is referred to as the head of the rule. An example of an access control policy from the CEI project is shown in Fig. 13.6a. Once encoded in SWRL through Protégé, this rule will be in the format shown in Fig. 13.6b. Here the arguments in the rule, such as ?req and ?course, are directly mapped to the related classes (model elements) in Protégé and their instances can be easily retrieved by a rule engine for interpretation.

a	b
$\forall r \in \text{ROLE}, \forall \text{req} \in \text{REQUEST}, \forall w \in \text{WORKFLOWSTATUS}, \forall op \in \text{OPERATIONWANT}, \forall \text{course} \in \text{COURSE}, \forall \text{format} \in \text{FORMAT}, \forall \text{agency} \in \text{AGENCY}, \forall loc \in \text{LOCATION}:$ <pre> role-permission(r,req,w,op) ← status(req,w) ∧ (1) bridging entity request(req, course, format, agency) ∧ (2) bridging entity location(agency, loc) ∧ (3) bridging entity inChargeLocation(r, loc) ∧ (4) geographical constraints inChargeCourse(r, course) ∧ (5) topic constraints inChargeFormat(r, format) (6) format constraints </pre>	$\boxed{\begin{array}{l} pl:rolePermission(?r,?p) \leftarrow \\ workflow(?w) \wedge \\ permissionRequest(?p,?req) \wedge \\ permissionOperation(?p,?op) \wedge \\ permissionStatus(?p,?w) \wedge \\ operationW(?op) \wedge \\ \end{array}}$ <pre> status (?req, ?w) ∧ (1) requestCourse (?req, ?course) ∧ requestFormat (?req, ?format) ∧ requestAgency (?req, ?agency) ∧ (2) agencyLocation (?agency, ?loc) ∧ (3) inChargeLocation (?r, ?loc) ∧ (4) inChargeCourse (?r, ?course) ∧ (5) inChargeFormat (?r, ?format) (6) </pre>

Fig. 13.6 An example of access policy for CEI (reprinted from [49]); (a) Access policy in first-order predicate logic, (b) Access policy in Protege SWRL

13.5.3 Interpretation of Access Control Policies

As described in Sect. 13.5.2, access control policies are encoded as SWRL rules. To interpret these policies, we use the jess engine for rule execution. jess is a rule engine and scripting environment written in the Java programming language [23]. It is small, light-weight, and fast. jess is a mature system that has been used to develop a broad range of applications [42, 59, 85]. It has already been integrated with Protégé and SWRL [18].

For a specific SWRL rule, the jess engine can map its arguments directly to the corresponding Protégé classes, retrieve the instances under these classes, apply the logic of the SWRL rule, and generate new instances (typically user-role assignments and role-permission assignments) as the consequence of the execution. With these instances of user-role assignments and role-permission assignments, we can judge whether a specific user should be assigned to a particular role; we can also decide whether a specific role should have access to a particular object at certain stages of workflow with a distinctive operation. This entire process of argument mapping, rule interpretation, and new instance generation is built together as Protégé add-ons.

A limitation of SWRL and jess is that they only support monotonic reasoning and hence cannot delete existing instances in Protégé. This presents a challenge to interpretation of access control policies when access revocation is required (and thus need to delete specific instances of role-permission assignments). To address this issue, we have developed a separate Protégé add-on that is specifically used for access revocation. This Protégé add-on searches and removes all relevant instances when interpreting a rule to revoke access permissions.

13.5.4 Application Layer

The application layer is where the access control policies are applied to specific cases and scenarios to make decisions on granting or denying access. This layer includes four major components: (1) a *policy enforcement* module, (2) an *access portal*, (3) an *object portal*, and (4) an *object storage* (see Fig. 13.4).

The access portal provides an interface for system users in particular roles. The policy enforcement module is linked with the policy interpretation layer to make decisions to grant or to deny access based on users, roles, objects, workflow statuses, operations, and constraints. The object portal provides a mechanism to generate new objects and to retire old objects. Once an object is generated, it stays in the object storage, where the policy enforcement module can search based on criteria defined in universal constraints. It is important to note that the specific functions and procedures defined at this layer may differ from application to application.

To apply the enhanced RBAC model to CEI for information access management, we mapped the components of the enhanced RBAC to specific entities in the CEI program. These mappings include: (1) training session \longleftrightarrow object, (2) training center \longleftrightarrow role, (3) employee of a training center \longleftrightarrow user, and (4) stage of a training session \longleftrightarrow workflow status.

Through the access portal and object portal, we imported the CEI program data for the period between April 1, 2011 and June 30, 2011, including all 409 healthcare organizations, 104 training sessions, 17 users (employees), 6 roles (training centers), and 44 user-role assignments. These data are transformed as Protégé instances. Through the policy enforcement module and the execution of the universal constraints encoded as SWRL rules, we can generate all role-permission assignments that define the level of access to specific training sessions by particular training centers in certain stages of training workflow. We report the findings from an evaluation study in Sect. 13.6.

13.5.5 Demonstration Tool

We have developed a demonstration tool for two purposes: (1) to provide a user interface to support the management of information access; and (2) to present access permissions under specific combinations of users, roles, objects, operations, and workflow statuses.

We implemented this tool as a Java application, with a screenshot shown in Fig. 13.7. The user interface of the demo tool includes four portions:

Portion A: Lists of all available users, roles, objects, and workflow statuses.

Portion B: Selections of specific users, roles, objects, and workflow statuses for examinations of information access management.

Portion C: Execution results after applying the access control policies defined in the enhanced RBAC to the selected users, roles, objects, and workflow statuses.

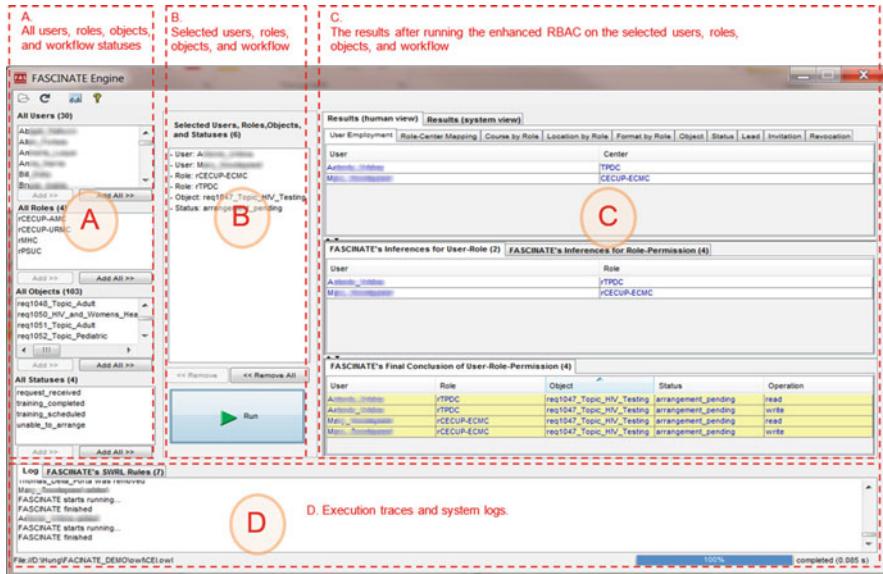


Fig. 13.7 A screenshot of the demo tool showing CEI access management (reprinted from [49])

Portion D: Execution traces and system logs.

Items in Portion B are a subset selected from Portion A. Once the items in Portion B are selected, we can click the “run” button to generate the results in Portion C. These results include: (1) all validated predicates (relations among users, roles, objects, operations, workflow statuses, and other entities) defined in the system; (2) all validated user-role mappings and role-permission mappings; and (3) assignments of permissions to users under specific roles at particular workflow stages. With these functions, the demo tool can be effectively used to examine access permissions in specific cases and scenarios.

13.6 Evaluation of the Enhanced RBAC Model

To examine the effectiveness of the enhanced RBAC model, we describe in this section an evaluation study to apply the model to CEI in order to manage information access in a specific context of collaborative processes. This study provides a first set of quantitative measures on effectiveness of the model, using the CEI application as a specific example. As shown in Sects. 13.4 and 13.5, the CEI project has complex requirements for information access management involving both team collaboration and workflow. It also has a rich set of real-world study cases that could be used for the evaluation. The CEI project already had in place a system (CEIAdmin) to manage information access in an ad hoc approach that could

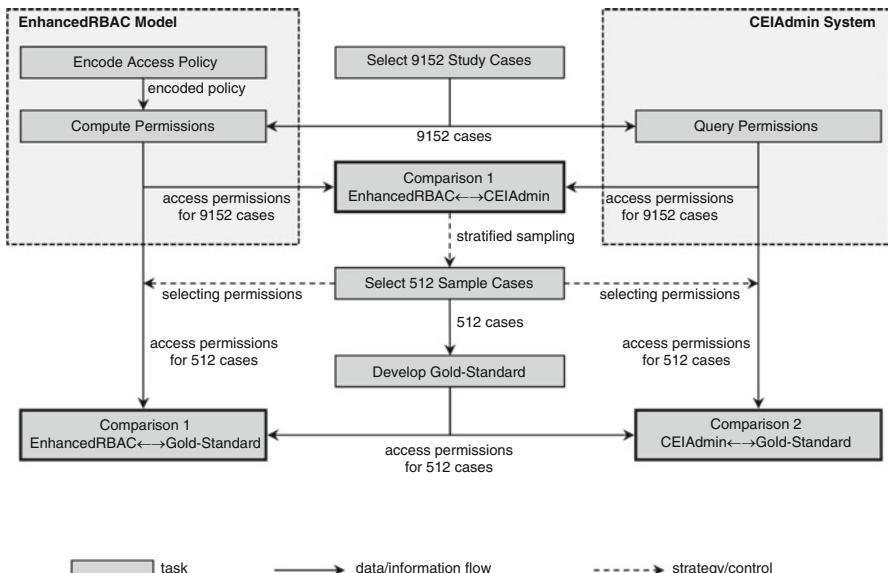


Fig. 13.8 Overall design of the evaluation study (reprinted from [52], with permission from Elsevier)

be used as a control. Therefore, using CEI as an application to evaluate the enhanced RBAC model can provide important insights on its effectiveness.

The overall study design consisted of two sets of comparisons: (1) the access permissions generated by the enhanced RBAC model vs. those generated by CEIAdmin, and (2) the access permissions generated by the enhanced RBAC model vs. a gold-standard. By performing the comparison in (1), we measured the degree of agreement between these two systems. For comparison (2), we first built a gold-standard through organizing an expert panel to manually review a selected set of study cases in order to determine the ideal access permissions (the ground truth) for each of them. Using these sample cases with the gold-standard access permissions, we then measured the effectiveness of the enhanced RBAC model (as well as CEIAdmin). The overall design of the evaluation study is shown in Fig. 13.8.

13.6.1 Selection of Study Cases

To select the study cases, we queried the CEI database on July 1, 2011 and obtained all onsite training sessions requested between April 1, 2011 and June 30, 2011. As the CEI project had a quarterly schedule for system upgrades, selection of study cases from this time period ensured a large enough sample size and, meanwhile, reduced potential biases due to the different versions of the CEIAdmin system.

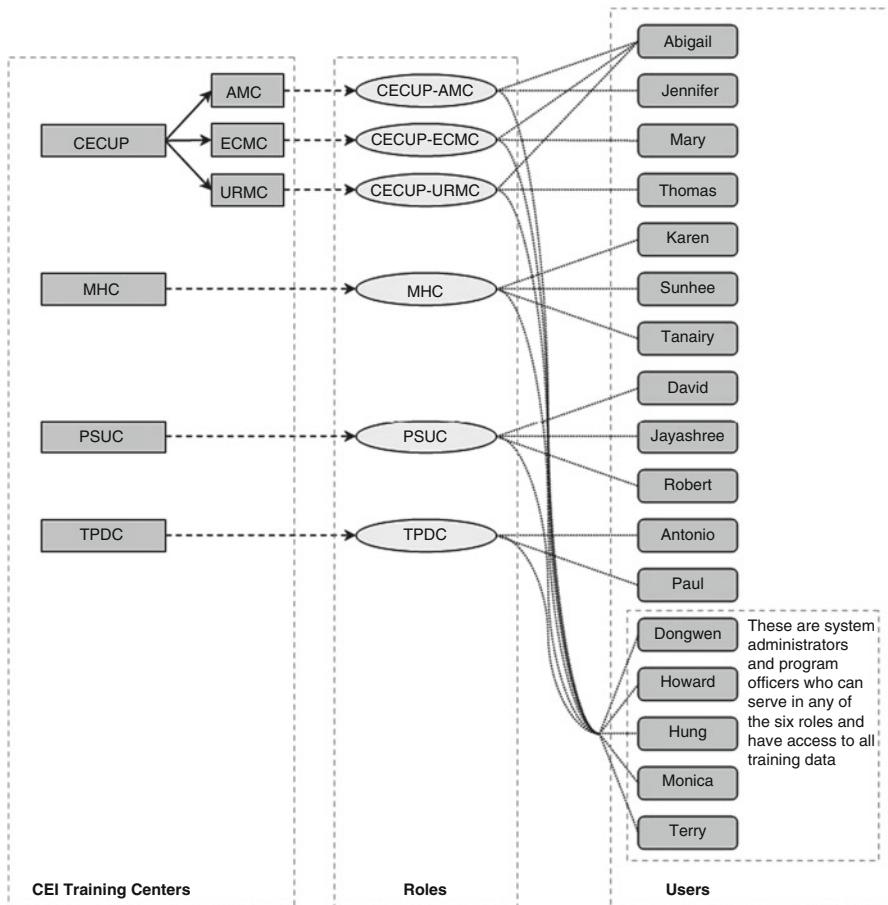


Fig. 13.9 Mappings of CEI Centers, system roles, and users (reprinted from [52], with permission from Elsevier)

The query generated a total of 104 training sessions requested by 38 healthcare organizations on 27 training topics in 5 workflow statuses. At the time when the query was made, CEIAdmin had 17 users, 6 roles, and a total of 44 user-role assignments, as shown in Fig. 13.9. For each training session, we defined 2 types of operations, i.e., “read” (review training data) and “write” (document training information and/or invite other CEI Centers for collaboration) [51]. By combining these 104 training sessions, 44 user-role assignments, and 2 operations, we created a total of 9152 ($104 \times 44 \times 2$) study cases. The profile of these study cases is shown in Table 13.2. It is important to note that this is a cross-sectional study, as the study cases were obtained through a system query at a specific time point.

Table 13.2 The profile of the study cases and the selected sample to build the gold-standard (reprinted from [52], with permission from Elsevier)

Characteristics of Study Cases	Total Cases	Sample Cases	
<i>User</i>			
Abigail	624	7 %	34
Antonio	208	2 %	12
David	208	2 %	10
Dongwen	1248	14 %	70
Howard	1248	14 %	66
Hung	1248	14 %	72
Jayashree	208	2 %	10
Jennifer	208	2 %	12
Karen	208	2 %	12
Mary	208	2 %	10
Monica	1248	14 %	68
Paul	208	2 %	14
Robert	208	2 %	14
Sunhee	208	2 %	12
Tanairy	208	2 %	14
Terry	1248	14 %	70
Thomas	208	2 %	12
Total	9152	100 %	512
<i>Role</i>			
CECUP—AMC	1456	16 %	74
CECUP—ECMC	1456	16 %	54
CECUP—URMC	1456	16 %	122
Mental Health Center	1664	18 %	108
Prevention and Substance Use Center	1664	18 %	76
Testing, Post-Exposure Prophylaxis, and Diagnosis Center	1456	16 %	78
Total	9152	100 %	512
<i>Course/Topic</i>			
Acute HIV infection—new frontiers for HIV prevention	176	2 %	16
Antiretroviral therapy—1 hour	88	1 %	20
Care of the triply diagnosed	176	2 %	10
Cognitive behavioral therapies in primary care settings	88	1 %	20
Cross cultural issues and cultural proficiency in ...	88	1 %	4
Hepatitis C and HIV co-infection	88	1 %	2
HIV and mental health	88	1 %	6
HIV and women's health	88	1 %	2
HIV testing	440	5 %	30
HIV testing for people with mental illness	88	1 %	2
Legal ethical issues	88	1 %	6
Management of alcohol use in HIV patients	88	1 %	8
Motivational interviewing in primary care settings	176	2 %	38
Opioid overdose prevention—role of naloxone in ...	88	1 %	6

(continued)

Table 13.2 (continued)

Characteristics of Study Cases	Total Cases	Sample Cases	
Other clinical psychiatric aspects of primary care	352	4 %	32 6 %
Post exposure prophylaxis	264	3 %	8 2 %
Resistance testing	88	1 %	2 0 %
The role of the primary care clinician in HIV care	352	4 %	0 0 %
Topic—adult	4488	49 %	208 41 %
Topic—ethical legal issues	88	1 %	6 1 %
Topic—HIV testing	704	8 %	30 6 %
Topic—mental health	176	2 %	8 2 %
Topic—pediatric	176	2 %	12 2 %
Topic—PEP	88	1 %	12 2 %
Topic—pharmacy	88	1 %	4 1 %
Topic—prevention	88	1 %	0 0 %
Topic—substance use	352	4 %	20 4 %
Total	9152	100 %	512 100 %
<i>Geographical Location</i>			
Metropolitan New York City area	1848	20 %	82 16 %
Upstate New York region	7304	80 %	430 84 %
Total	9152	100 %	512 100 %
<i>Workflow Status</i>			
Request received	880	10 %	30 6 %
Arrangement pending	528	6 %	88 17 %
Training scheduled	5896	64 %	342 67 %
Training completed	1760	19 %	48 9 %
Unable to arrange	88	1 %	4 1 %
Total	9152	100 %	512 100 %
<i>Operation</i>			
Read	4576	50 %	256 50 %
Write	4576	50 %	256 50 %
Total	9152	100 %	512 100 %

13.6.2 Access Permissions Computed with the Enhanced RBAC Model and the CEIAdmin System

To compute access permissions with the enhanced RBAC model, we first used the Protégé tool [71] to encode the model, with the universal constraints defined in SWRL [66]. We then imported the 9152 study cases as ontology instances. We leveraged an existing Protégé add-on with an external Jess package [23] to interpret the encoded constraints and to determine whether access permissions should be granted to specific study cases. By the end of this process, we were able to make a

decision based on the enhanced RBAC model with each study case either to grant or to deny access. The technical details of the implementation can be found in Sect. 13.5.

For the CEIAdmin system, if a specific user in a particular role logged in, the hard-coded logic would be triggered and the training sessions to which this user had access should be presented on the screen [51]. Thus, by capturing the data on the screen for each user-role assignment, we were able to obtain all the study cases with which access were granted to this specific user-role pair. Once these positive cases (access granted) were identified, we sifted them out from the entire case pool to identify the remaining (negative) cases, with which access permissions were denied by the CEIAdmin.

13.6.3 Comparison Between the Enhanced RBAC Model and the CEIAdmin System

Since the CEIAdmin system has already been in production use, we can assume it is reasonably good to determine access permissions for specific training sessions. Thus, by comparing the results computed from the enhanced RBAC model and CEIAdmin, we can indirectly measure the effectiveness of the enhanced RBAC using CEIAdmin as a benchmark. For this purpose, we compared the two systems based on: (1) only the “read” operation, (2) only the “write” operation, and (3) both the “read” and “write” operations. For each comparison, we calculated the kappa value to measure the degree of agreement [24]. An interesting phenomenon for the CEI project is that the “write” permission is embedded within the “read” permission (i.e., a role always has the “read” permission if it already has the “write” permission). Therefore, the access permission for a specific training session can be redefined as three mutually exclusive outcomes: “no” access, “read-only” access, and “read+write” access. To measure the degree of agreement based on this formulation of outcomes, we transformed the results from the previous computation and recalculated the kappa value. This analysis with the triple outcomes provided a more accurate comparison between the enhanced RBAC and CEIAdmin.

13.6.4 Development of the Gold-Standard

Simply comparing the enhanced RBAC model with the CEIAdmin system cannot provide a complete measurement on its effectiveness. For example, even if the two systems agree with each other on a specific case, it is still possible that both systems are wrong (although this chance is low if we assume that the effectiveness of CEIAdmin is reasonably good). In order to have an objective measurement on

system effectiveness, we need to have a gold-standard with correct answers on access permissions for a specific study case. Following the common practice in developing gold-standards for evaluation of information systems [24], we organized a panel with four domain experts, who participated in design of the CEIAdmin system and were directly involved in daily management of the CEI project, to review study cases and to develop standard (correct) answers. We then used the standard answer for each case as the reference to measure the effectiveness of the enhanced RBAC model and the CEIAdmin system.

Since it was impractical to manually review all the 9152 study cases, we decided to first obtain a representative sample from the case pool and then to build the gold-standard based on this sample. The results from the previous comparisons between the enhanced RBAC and CEIAdmin provided important information to guide the sampling process. Specifically, for those “matched” study cases (i.e., the enhanced RBAC and CEIAdmin had a consistent decision on granting or denying access), the chances that both systems were wrong should be relatively low; for those “unmatched” study cases (i.e., the enhanced RBAC and CEIAdmin had different decisions on granting or denying access), at least one of the systems should be wrong. Therefore, we would need to pay more attention to those “unmatched” cases when performing manual review to build the gold-standard.

For this reason, we decided to divide the study cases into the “matched” stratum and the “unmatched” stratum. We then used a stratified random sampling technique to sample the two strata separately, with a higher sampling rate for the “unmatched” stratum such that a larger proportion of the “unmatched” cases could be manually reviewed. We determined a sampling rate of 5 % for the “matched” stratum and a sampling rate of 20 % for the “unmatched” stratum, with a total of 278 study cases obtained. In this way we mitigated the imbalance between the two strata, and meanwhile controlled the total number of sample cases so to ensure the feasibility to manually review all of them.

As we planned to use an expert panel to review the sample cases and to develop the gold-standard, it would be more intuitive and efficient for the expert judges to review the “read” and “write” permissions together for a specific training request, rather than asking them to check only the “read” or only the “write” access (we called the two study cases “complementary” to each other if they had the same user, role, training session, and workflow status but different access operations). Thus, we decided to include into the sample additional study cases if their complementary cases had already been selected. By the end of the sampling process, we had a total of 512 study cases. The profile of this sample is shown in Table 13.2.

To control potential biases, we used a partial factorial design and a blocked randomization technique to assign study cases to each judge such that: (1) a study case was randomly assigned to a judge, (2) each case was reviewed by two judges, and (3) each judge reviewed half of the sample cases. For each study case in the sample, a judge first determined whether access should be granted; if so, he/she needed to decide whether it should be a “read-only” access or a

Table 13.3 Assignment of sample cases to judges and results from the first round review^a

	DW	MB	TD	XHL	Total
DW	–	90 (79)	88 (87)	82 (81)	260 (247)
MB	90 (79)	–	88 (77)	82 (82)	260 (238)
TD	88 (87)	88 (77)	–	82 (81)	258 (245)
XHL	82 (81)	82 (82)	82 (81)	–	246 (244)
Total	260 (247)	260 (238)	258 (245)	246 (244)	1024 (974)

^aDW, MB, TD, and XHL are the judges. For each cell, the first number is the cases assigned to the two judges and the second number in parenthesis is the cases with consistent results from the first round of review by the two judges. Table reprinted from [52], with permission from Elsevier

“read+write” access. Each judge independently made these decisions for all study cases assigned to him/her. During this process, all judges except the one who managed the study data were blinded to the execution results generated by the enhanced RBAC and CEIAdmin. Out of the 1024 (512×2) cases, the judges had achieved initial consensus on 974 (95 %). For the remaining 50 cases with discrepancy in judge opinions, the related judges sit down together, reviewed the cases, and made new decisions until consensus were reached. The assignment of sample cases to judges and the results from the first round review are presented in Table 13.3. Figure 13.10 is a screenshot of the online tool that was used by the judges in order to make decisions on specific study cases to build the gold-standard.

13.6.5 Measuring Effectiveness Based on Gold-Standard

To measure the effectiveness of the enhanced RBAC model when applied to the sample cases, we first formulated a 3×3 table to compare it with the gold-standard based on the previously described three outcomes (“no” access, “read-only” access, and “read+write” access).

By cutting the results between the “no” access and the other two outcomes, we were able to evaluate the effectiveness of the enhanced RBAC model on “read” operation; by cutting the results between the “read+write” access and the other two outcomes, we were able to evaluate the effectiveness of the enhanced RBAC model on “write” operation. By adding up the outcomes for both the “read” and “write” operations, we were able to evaluate the overall effectiveness of the enhanced RBAC model.

For each comparison, we calculated sensitivity, specificity, and accuracy as the specific measures. Here sensitivity is the proportion of actual positives (access granted as indicated by the gold-standard) that are correctly identified by the enhanced RBAC model; specificity is the proportion of actual negatives (access

Fig. 13.10 A screenshot of the online tool used by the judges to build the gold-standard (reprinted from [52], with permission from Elsevier)

You are logged on as user: [Logout](#)

Change username

GOLD STANDARD VERIFICATION

You have completed 138 of total 138 cases. [Show All](#)

CaseID	CEI Staff	CEI Center (Staff's Role)	ReqID	Requesting Agency	Agency's County	Requested Course/Topic	Request Status
115334	Antonio Urbina	TPDC	1007	University of Rochester Medical Center/Brown Memorial Hospital	Morrie	Topic: Adult	TEF

Questions:

1) Can Antonio Urbina (with role: TPDC) access this request?
 Yes No
Initiated by: CECUP-URMC

2) What access permission does he/she has?
 Read-only Read & Write (Documentation)
Initiated by: CECUP-URMC

[Previous](#) [Next](#)

Hide Policy

Notes

1. A CEI staff may have multiple roles. The staff roles included here are: CECUP-AMC, CECUP-ECMC, CECUP-URMC, MHC, PSUC, and TPDC.
2. Request status include: New Request, Call Back, Scheduling, and TEF.
3. All changes to the form are automatically saved.
4. Please click "View" in "ReqID" cell to check all the course/topics in this request.

Policies

1. Geographical Responsibilities
 - a. CECUP-AMC, CECUP-URMC, and CECUP-ECMC each is responsible for a catchment area (see 'CEI Center and County Mapping').
 - b. CECUP-AMC and TPDC are responsible for requests from all counties.
2. Course/Topic Responsibilities
 - a. MHC, PSUC, and TPDC each is responsible for a list of courses/topics (see 'CEI Center and Course/Topic Mapping').
 - b. CECUP-AMC, CECUP-URMC, and CECUP-ECMC, are responsible for all courses/topics.
3. Documenting Lead Center
 - a. For a request in status 'Scheduling' or 'TEF', only one CEI Center can be the Lead Center (i.e. has 'write' permission).
 - b. If a request is in status 'New Request' or 'Call Back', any collaborating center can be the documenting center (i.e. has 'write' permission).
4. Initiation
 - a. If Center A is the Lead Center for a request in status 'Scheduling' or 'TEF', if Center B initiates another Center B to collaborate.
 - b. If Center A invites Center B for collaboration, Center A grants Center B permission to initiate (i.e. Center B can initiate request).
 - c. If Center A invites Center B for collaboration and assigns Center B as the Lead Center, Center A immediately releases itself as the Lead Center. After that, Center B will become the Lead Center, and Center A will become a collaborating center with 'read' permission.
5. If a request includes multiple courses/topics, a CEI Center involved with one course/topic will have 'read' permission to other courses/topics of the same request. See example.

CEI Center and Course/Topic Mapping

CECUP-AMC, CECUP-URMC, CECUP-ECMC

- All Courses/Topics

MHC

- Course: Care of the Trauma Diagnosed
- Course: Cognitive-Behavioral Therapies in Primary Care Settings
- Course: Cross-Cultural Issues & Cultural Proficiency in Psychiatric Diagnosis & Treatment
- Course: Diagnostic Assessment & Treatment Recommendations
- Course: Evidence-Based Practice
- Course: HIV Testing for People with Mental Illness
- Course: Mental Health Medication Evaluation
- Course: Mental Health Legal Environment
- Course: Motivational Interviewing in Primary Care Settings
- Course: Neuropsychiatric Aspects of HIV Infection
- Course: Other Clinical/Practical Aspects of Primary Care
- Topic: Mental Health

PSUC

- Course: Adherence to ART among Substance Users
- Course: Best Practices for Continuity of Care for HIV+ Woman
- Course: Clinical Management of Substance Abuse
- Course: Management of Alcohol Use in HIV Patients
- Course: Opioid Overdose Prevention: Role of Naloxone in the Community
- Course: Prevention of Substance Use in HIV+ Women
- Course: Screening & Assessment for Substance Using HIV Infected Patients
- Course: Smoking Cessation in HIV Infected Patients
- Course: Substance Use Treatment Modalities for HIV+ Substance Users
- Topic: Prevention
- Topic: Substance Use

TPDC

- Course: Acute HIV Infection: New Frontiers for HIV Prevention
- Course: HIV Testing
- Course: HIV/AIDS Prophylaxis
- Topic: Acute HIV Infection
- Topic: HIV Testing
- Topic: PEP

CEI Center and County Mapping

CECUP-AMC	CECUP-ECMC	CECUP-URMC
<ul style="list-style-type: none"> • Albany • Greene • Herkimer • Madison • Montgomery • Onondaga • Orange • Otsego • Oswego • Putnam • Rensselaer • Saratoga • Schenectady • Schuyler • Sullivan • Ulster • Warren • Washington 	<ul style="list-style-type: none"> • Allegany • Chautauqua • Erie • Genesee • Niagara • Orleans • Wyoming 	<ul style="list-style-type: none"> • Cayuga • Chenango • Cortland • Jefferson • Lewis • Livingston • Monroe • Ontario • Oneida • Seneca • Steuben • Tioga • Tompkins • Wayne • Yates
• All Counties		

denied as indicated by the gold-standard) that are correctly identified by the enhanced RBAC model; accuracy is the proportion of actual positives or negatives (as indicated by the gold-standard) that are correctly identified by the enhanced RBAC model. As a comparison, we performed the same set of measurements on CEIAdmin. For all analyses in this study, we used the SPSS statistical package [36]. We report the detailed results in the next section.

13.6.6 Results

When formulated with three outcomes (“no” access, “read-only” access, and “read+write” access), the enhanced RBAC model and the CEIAdmin system agreed on 4230 out of the 4576 study cases. With a kappa value of 0.80 (95 % CI: 0.78–0.82), these two systems demonstrated a high level of agreement. When formulated with two outcomes (granting or denying access), the two systems agreed on 4399 cases for the “read” operation ($\kappa = 0.89$, 95 % CI: 0.88–0.91) and 4400 cases for the “write” operation ($\kappa = 0.88$, 95 % CI: 0.86–0.90). Combining both, the two systems agreed on 8799 out of the total 9152 cases ($\kappa = 0.89$, 95 % CI: 0.88–0.90). These comparisons have shown that the enhanced RBAC model has achieved a high level of agreement with CEIAdmin. The detailed results are reported in Table 13.4.

When evaluated against the gold-standard, the enhanced RBAC model had a correct answer for 251 out of the 256 cases when the results were formulated with three outcomes (accuracy = 98 %, 95 % CI: 97–100 %). When transformed into two outcomes and measured by the “read” operation, the enhanced RBAC model achieved a sensitivity of 97 % (95 % CI: 94–99 %), a specificity of 100 % (95 % CI: 100–100 %), and an accuracy of 98 % (95 % CI: 96–100 %). Based on two outcomes and measured by the “write” operation, the enhanced RBAC model achieved a sensitivity of 100 % (95 % CI: 100–100 %), a specificity of 100 % (95 % CI: 100–100 %), and an accuracy of 100 % (95 % CI: 100–100 %). Combining both, the enhanced RBAC model achieved a sensitivity of 98 % (95 % CI: 96–100 %), a specificity of 100 % (95 % CI: 100–100 %), and an accuracy of 99 % (95 % CI: 98–100 %). As a comparison, we performed the same set of measurements on CEIAdmin. The results have shown that the CEIAdmin system had an overall accuracy of 76 % (95 % CI: 70–81 %) when the results were formulated with three outcomes. It achieved sensitivities at the level of 100 %, specificities in the range of 61–97 %, and accuracies in the range of 76–99 % when the results were formulated with two outcomes. A complete report of the results is shown in Table 13.5.

Table 13.4 Comparison between the enhanced RBAC model and the CEIAdmin system

Comparison with three outcomes kappa = 0.80 (95 % CI: 0.78–0.82)		CEIAdmin system			
		Read + write	Read only	No	Total
Enhanced RBAC model	Read + write	820	0	0	820
	Read only	169	15	0	184
	No	7	170	3395	3572
	Total	996	185	3395	4576
Comparison with two outcomes “read” operation kappa = 0.89 (95 % CI: 0.88–0.91)		CEIAdmin system			
		Yes	No	Total	
Enhanced RBAC model	Yes	1004		0	1004
	No	177		3395	3572
	Total	1181		3395	4576
Comparison with two outcomes “write” operation kappa = 0.88 (95 % CI: 0.86–0.90)		CEIAdmin system			
		Yes	No	Total	
Enhanced RBAC model	Yes	820		0	820
	No	176		3580	3756
	Total	996		3580	4576
Comparison with two outcomes all operations kappa = 0.89 (95 % CI: 0.88–0.90)		CEIAdmin system			
		Yes	No	Total	
Enhanced RBAC model	Yes	1824		0	1824
	No	353		6975	7328
	Total	2177		6975	9152

D: we performed discrepancy analyses based on each category of cases (see Sect. 13.7.3). Table reprinted from [52], with permission from Elsevier

13.7 Discussion

13.7.1 Features of the Enhanced RBAC Model

Access control has been applied to many applications in healthcare settings [7, 13, 28, 44, 50, 56, 60, 68, 75, 79]. Since confidentiality of clinical information is an essential requirement, most of the previous work focused on applications for patient care. In the context of clinical education, controlling access to specific information was not a primary concern by tradition. However, as we are moving toward a system-wide and team-based approach to providing clinical education [78, 84], more and more training programs are involving multiple collaborative parties and complex training workflows. The CEI project is a perfect example. To our knowledge, this work is the first application of an information access control model to clinical education. The case study on the CEI project has provided us important insights in designing an information access control model that can be generalized to various applications for clinical education, biomedical research, and patient care, which will be the direction for our future research.

Table 13.5 Measuring the effectiveness of the enhanced RBAC model and the CEIAdmin system with a gold-standard

		Gold-standard				
		Read + write	0	Read only	No	Total
Measurement with three outcomes Enhanced RBAC model accuracy = 98% (95% CI: 97–100%)		Read + write 98	0	0	0	98
Read only		0	63	0	0	63
No		0	^A 5	90	95	95
Total		98	68	90	256	
CEIAdmin system accuracy = 76% (95% CI: 70–81%)		Read + write 98	C59	B3	160	
Read only		0	9	0	9	
No		0	0	87	87	87
Total		98	68	90	256	
Measurement with two outcomes—“read” operation		Gold-standard				
		Yes		No	Total	
Enhanced RBAC model Sensitivity = 97% (94–99%), specificity = 100% (100–100%), accuracy = 98% (96–100%)		Yes 161		0	161	
No			5	90	95	95
Total		166		90	256	
CEIAdmin system Sensitivity = 100% (100–100%), specificity = 97% (92–100%), accuracy = 99% (97–100%)		Yes 166		3	169	
No			0	87	87	87
Total		166		90	256	

(continued)

Table 13.5 (continued)

		Gold-standard					
		Read + write	Read only	No	Total		
		Gold-standard					
Measurement with three outcomes							
Comparison with two outcomes—"write" operation		Yes	98	0	0	98	
Enhanced RBAC model	Sensitivity = 100 % (100–100 %), specificity = 100 % (100–100 %), accuracy = 100 % (100–100 %)	No	0	158	158		
	Total	98		158	256		
CEIAadmin system	Sensitivity = 100 % (100–100 %), specificity = 61 % (53–68 %), accuracy = 76 % (71–81 %)	Yes	98	0	62	160	
	No	0		96	96		
	Total	98		158	256		
Measurement with two outcomes—all operations							
Enhanced RBAC model	Sensitivity = 98 % (96–100 %), specificity = 100 % (100–100 %), accuracy = 99 % (98–100 %)	Yes	259	0	0	259	
	No	5		248	253		
	Total	264		248	512		
CEIAadmin system	Sensitivity = 100 % (100–100 %), specificity = 74 % (68–79 %), accuracy = 87 % (85–90 %)	Yes	264	0	65	329	
	No	0		183	183		
	Total	264		248	512		

A,B,C: we performed error analyses based on each category of cases (see Sect. 13.7.3). In parenthesis are percentages computed based on a 95 % Confidence Interval (CI). Table reprinted from [52], with permission from Elsevier

Access control is typically utilized to protect certain information, and thus is more frequently implemented to limit information access. In the case of CEI, facilitation of team coordination and collaboration was a primary goal. Therefore we used access control not only to limit access to information (i.e., to sift out requests beyond the catchment area of a specific CEI Center) but also to enable access to information to facilitate collaboration (i.e., to review the progress of training workflow documented by a collaborating CEI Center). Implementing access management in the context of team coordination and collaboration for both facilitation and control of information access is another contribution of this work.

To manage information access in a collaborative environment, we identified unique concepts, such as bridging entity and contributing attribute, and incorporated them into the model as specific types of universal constraints. From the CEI project, we have developed geographical constraints, training topic constraints, and training format constraints to model specific types of contributing attributes. Obviously additional types of universal constraints can and will be identified from other applications. Certain types of universal constraints (such as the ones we identified from the CEI project) can be applied to similar applications. As a long term goal, we plan to assemble the constraint sets identified from various application scenarios to formulate a UNIversal Constraint ONtology (UNICON), which can be utilized as a knowledge base to drive information access management in a variety of situations.

To incorporate workflow context into the modeling of access control, we selected to expand the core RBAC definition of access permission to include workflow status. Consequently, the enhanced RBAC is expressive in defining the workflow context for access permission. For example, the CEI project requires that access delegation can only happen in later stages of the workflow (“*training-scheduled*” and “*training-completed*”). This workflow context-specific requirement on access management can be easily specified with the enhanced RBAC model. On the other hand, including workflow status as another dimension in the definition of access permission introduces complexity in the specification of permissions and constraints. A potential solution to balance the expressiveness and simplicity is to allow the access permissions to carry over forward by default when the workflow is changing, and meanwhile using new constraints to overwrite the default permissions when necessary. We used this concept to define the access permissions for the CEI project. Future research needs to focus on identifying general methods or strategies to address this challenge.

The access delegation in the enhanced RBAC model is unique in that it targets on access permissions for only specific objects. This feature provides flexibility to enable information sharing in a collaborative environment and meanwhile limits the access to only those objects that are relevant to the collaboration. Thus, it has balanced nicely between flexibility and need-based access. Previous research of access delegation under RBAC focused on study of the relationship between users and roles [90, 91]. Our work for the first time modelled the relationship between roles and objects in access delegation. For example, when CECUP invites MHC to collaborate on training request “1090”, access delegation is limited to only this specific request; in comparison, other methods will require to grant MHC access

permission to all CECUP's data. Therefore, the enhanced RBAC can achieve better security and effectiveness in information access management. It is important to note that an access control model only provides a means to define access policies. In order to take full advantage of the enhanced RBAC model, we need to leverage its capacity and configure a set of access policies that fit best with the domain requirements.

13.7.2 System Framework for Implementation

Several layered approaches to implement access control policies have been reported in the literature [26, 50, 65, 67]. For example, the implementation of the eXtensible Access Control Markup Language (XACML) [65] was based on a series of function modules/points such as Policy Administration Point (PAP), Policy Decision Point (PDP), Policy Enforcement Point (PEP), and Policy Information Point (PIP). However, the early versions of XACML did not provide native support to RBAC [21]. Even the specialized XACML profiles were not able to support many relevant constraints. To address this issue, there was a previous exploration to introduce XACML + OWL, a framework that integrates OWL ontologies and XACML policies, to support RBAC.

In our implementation of the enhanced RBAC, the system framework and its functional layers can be mapped to many functional modules in XACML implementation. Specific mappings include: the Protégé environment in the enhanced RBAC is functioning as PAP; the jess engine and revocation add-on are functioning as PDP; the policy enforcement module in Application Layer is functioning as PEP; and the Access Portal and Object Portal in Application Layer are functioning as PIP.

Beyond the standard XACML implementation, we have defined unique structures in the enhanced RBAC system framework. These structures include: (1) an external workflow engine to manage workflow context; (2) a three-level schema for definition of access control policies to differentiate the core RBAC, the extension of RBAC, and the policy instances; and (3) specific entities and relations defined for particular applications, which are processed through the Context Handler.

The mappings between the functional components of the enhanced RBAC system framework and the standard XACML implementation are shown in Fig. 13.11. Here the integration with an external workflow engine in (1) can support access management in specific context of workflow, which is an enhancement of the core RBAC model. The three-level schema for definition of access control policy in (2) can facilitate the continuous development of the enhanced RBAC model. The application-specific entities and relations defined in (3) can custom-tailor the system framework to particular domain applications for effective implementation of access control policies. These features define the unique contributions of this work.

In the prototype implementation of the enhanced RBAC model for the CEI project, we enumerated all possibilities for a request (training topic, agency location, training format, and workflow status) when defining the access control policies.

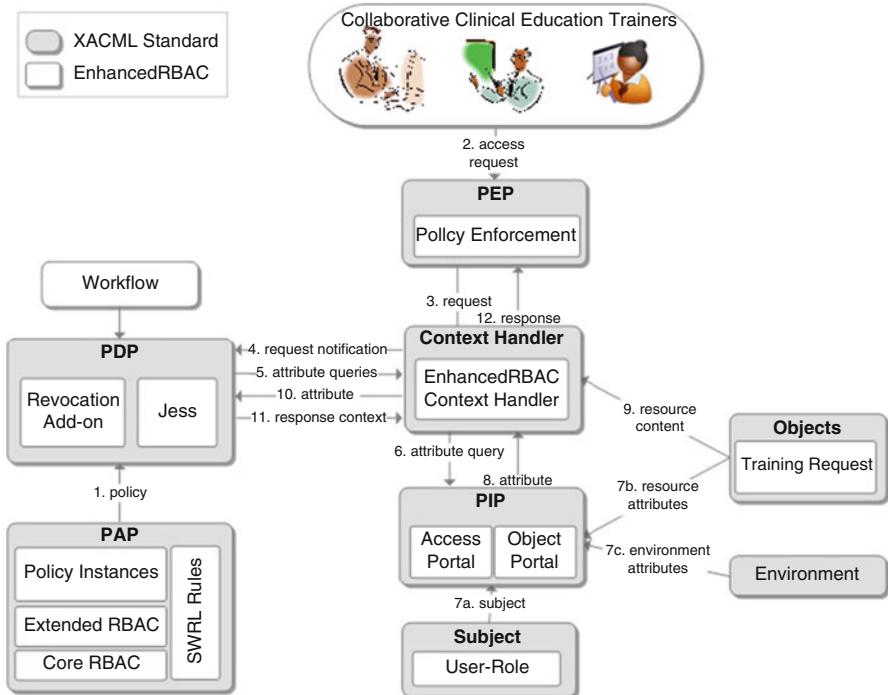


Fig. 13.11 Mapping the enhanced RBAC framework to XACML (reprinted from [49])

Therefore, for any request there was always a policy applicable. In this way, we guaranteed the completeness of the policies. In addition, we followed the convention to define policies only to grant access, with a default behavior to deny access if no policy was applicable. Meanwhile, we implemented a separate function to remove a previously granted access after it was revoked [49]. In this way, we ensured the soundness of the policies. Currently, we do not have tools to automatically detect and resolve potential policy conflicts. Others have already developed a few techniques to address this issue [4, 9, 22, 82]. This could be another direction for our future exploration.

13.7.3 Evaluation

13.7.3.1 Overall Approach

The overall strategy of the conducted evaluation study was situated on two sets of measurements: (1) agreement between the enhanced RBAC model and the CEIAdmin system, and (2) effectiveness of the enhanced RBAC model when evaluated against a gold-standard.

For the first set of measurements, since they could be performed on results automatically generated by systems, we could afford to use a very large sample. In this study, we used all 9152 study cases obtained from a period of 3 months. For the second set of measurement, we were limited by the capacity to build a gold-standard (usually through expert panel review) [24] and therefore the sample size had to be smaller. In this study, we selected a sample of 512 cases, which was sufficient but the statistical power was not at the same level as the first set comparisons.

In terms of the potential issues identified from these measurements, the first set of comparisons could find out the differences between the two systems, while the second set of measurements could pinpoint the specific causes of these differences and recognize the problems that were common to both systems. By combining these two sets of measures, we were able to perform an effective and powerful evaluation. To our knowledge, this is the first evaluation study that has provided empirical evidence on effectiveness of an access control system through benchmark measures of sensitivity and specificity, and degree of agreements with a control system. We have used a similar strategy in study design for other projects and successfully performed evaluations on different information models [86, 87]. We believe that this method can be generalized to additional domains and applications.

13.7.3.2 Error Analyses

As shown in Table 13.5, neither the enhanced RBAC model nor the CEIAdmin system was perfect when evaluated against the gold-standard. To assess the problems identified from the evaluation, we performed error analyses on the cases in Category A, B, and C (see Table 13.5).

Here Category A was the false negative of the enhanced RBAC model (access wrongly denied when “read” should be granted); Category B was the false positive of the CEIAdmin system (“read+write” wrongly granted when access should be denied); and Category C was another type of error of CEIAdmin (“read+write” access wrongly granted when “read-only” should be granted). The results have shown that:

1. The error for the 5 cases in Category A was caused by a mistake in preparing invitation data for the study cases to feed the enhanced RBAC model.
2. The error for the 3 cases in Category B was caused by adding/deleting data directly by a system administrator to/from the CEIAdmin database, which interfered with CEIAdmin’s hard-coded logic for access management.
3. The error for the 59 cases in Category C was caused by an administrative decision in developing CEIAdmin to allow a CEI Center staff to retrospectively document call-back information in an earlier workflow status of a training session, which led to two different workflow statuses for the same training session under specific scenarios.

In addition to error analyses, we reviewed the discrepancies between the enhanced RBAC model and the CEIAdmin system (see Table 13.4) and identified another system error from the 170 cases in Category D, different from any of the previous

situations. This error was caused by a minor inconsistency in definition of objects (training requests) between the enhanced RBAC model and the CEIAdmin system. Specifically, when a healthcare organization requested multiple training sessions at the same time, CEIAdmin allowed staff from a specific CEI Center to have access to all related training sessions as long as this CEI Center was involved in collaborative training for one of them (Policy #5 in Fig. 13.10). When specifying the access policies for the enhanced RBAC model, however, we only defined a single training session as the basic unit of an access object, and thus the encoded policies did not cover the situation when multiple training sessions were bundled together.

As shown in the results, there were 62 error cases for CEIAdmin (3 in Category B and 59 in Category C) but only 5 for the enhanced RBAC model (in Category A). Therefore, the overall accuracy of the enhanced RBAC model was much higher than CEIAdmin (98 % vs. 76 %). It was interesting to note that a simple administrative decision (leading to Category C error of CEIAdmin) and a walk-around solution to address data issues (leading to Category B error of CEIAdmin) could significantly impact the effectiveness in access management. Here the error in Category A could be corrected through data reformulation. The error in Category B was a rare (but possible) situation of a real world application (such as CEIAdmin), which we would not be able to prevent completely. The error in Category C could be addressed by disallowing retrospective documentation or through a refinement of access object. The error in Category D could be resolved by including additional access policies into the enhanced RBAC model.

13.7.3.3 Qualitative Measures

Beyond the quantitative measures on effectiveness, there are many other aspects in the evaluation of an access control system. Here we discuss the enhanced RBAC model with regard to a selected set of related qualitative measures based on the guidelines proposed by Hu et al. [34, 35]. Specifically, the enhanced RBAC model supports the following features:

1. Auditing of access through logs of all granted and denied requests [49, 51].
2. Easy discovery of access privileges and capabilities with any combination of users, roles, objects, and workflow statuses through a demonstration tool [49].
3. Easy privilege assignments with the Protégé tool through policy encoding and role definition [49, 51, 71].
4. Syntactic and semantic specification of access control rules through the use of SWRL [49, 80].
5. Enforcement of least privilege principle through precisely defined access rights in collaborative workflow [51].
6. Separation of duty through roles, workflow contexts, universal constraints [51].
7. Resolving conflict of access policies through the external Jess package [37, 49].
8. Awareness of situation through workflow status and enforcement of access policies in these situations [51].

9. Standard expression of access policies through XACML [51].
10. Good system performance on response time, policy deposit/retrieval, and integration with user authentication [49].
11. Capacity of policy import and export through Protégé [49, 51, 71].
12. Availability of graphical user interface and system APIs for both the Protégé and the demonstration tool [49, 71].
13. Verification of compliance with access policies in specific scenarios through the enhanced RBAC system framework and the associated tools [49, 51].

Currently, the enhanced RBAC model does not support multiple granularities in definition of objects. This is an area we would like to explore as a next step.

13.7.4 Limitations

A major limitation in the development of the enhanced RBAC model is that it was derived from a specific application for clinical education. Its generalizability, therefore, needs to be further examined in other applications and scenarios. For example, when managing information access in collaborative clinical workflow, the information access might be unpredictable but the needs to access certain patient information are immediate. To address this requirement, we can apply the enhanced RBAC model to define a policy to grant access (of patient records) to all clinicians in the same clinical unit. Depending on application requirements, we can make this policy more stringent (for example, requiring only those clinicians on duty can have access); or we can relax this policy to allow inviting experts from another clinical unit to consult on a patient case through access delegation. Apparently, the effectiveness of these policies needs to be evaluated in real world applications in order to identify the best that fits with domain requirements.

In this chapter, we only discussed the enhanced RBAC model for anterior control of information access. Another category of its use is for posterior auditing of information systems to identify incompliances with access policies. Since auditing is a common practice in clinical information systems, this could potentially be an important application for the enhanced RBAC model.

The evaluation study was based on a cross-sectional design and we did not follow up individual training sessions through the training process. Performing such a longitudinal study would require integration of the enhanced RBAC model with a workflow engine, which could be a future direction for our research. Nevertheless, our cross-sectional study included cases in all five workflow statuses (see Table 13.2). We therefore believe that the enhanced RBAC model should still be valid for the entire training process if we perform a longitudinal study, although specific measurements are required to confirm this projection.

When developing the gold-standard for the evaluation, the investigators served as the judges, which might introduce assessment bias [24]. Since the judges should be the domain experts who understand the information resources under evaluation [24],

we could not completely prevent this possibility. To reduce potential assessment bias, we blinded the system execution results from three of the four judges, and required each study case to be independently reviewed by two judges until a consensus was reached. In addition, we used CEIAdmin, which was developed by the same group of investigators, as a control in the evaluation. We therefore believe that the potential assessment bias was, if not completely prevented, well under control.

13.8 Conclusion

We have enhanced the RBAC model through: (1) formulating universal constraints, (2) defining bridging entities and contributing attributes, (3) extending access permission to include workflow context, (4) synthesizing a role-based access delegation model to target on specific objects, and (5) developing domain ontologies as instantiations of the general model to specific applications.

Based on this enhanced RBAC model, we have developed a system framework to implement policies for information access management in collaborative processes. We have successfully applied the enhanced RBAC model and the system framework to the CEI project to address the specific needs on information access management in the combined context of team collaboration and workflow.

An initial evaluation through comparison with a control system has shown a high level of agreement. When evaluated against a gold-standard, the enhanced RBAC model has achieved a very good measure on sensitivity and a perfect score on specificity. These initial results indicate that the enhanced RBAC model can be effectively used to manage information access in collaborative processes.

Future research is required to incrementally develop additional types of universal constraints, to further investigate how the workflow context and access delegation can be enriched to support the various needs on information access management in collaborative processes, to extend the system framework to support the continuous development of the enhanced RBAC model, to perform longitudinal evaluation studies, and to examine the generalizability of the enhanced RBAC model for other applications in clinical education, biomedical research, and patient care.

Acknowledgements The CEI project is sponsored by New York State Department of Health AIDS Institute through Contracts #C024882 and #C023557. We would like to thank Amneris Luque, Monica Barbosa, Terry Doll, Matthew Bernhardt, and Thomas Della Porta for their contributions. We would like to thank CEI program staff Howard Lavigne, Cheryl Smith, Lyn Stevens, Bruce Agins and the colleagues from the other CEI Centers for their support.

References

1. Alam, M., Zhang, X., Khan, K., Ali, G.: xDAuth: a scalable and lightweight framework for cross domain access control and delegation. In: Proceedings of the 16th ACM Symposium on Access Control Models and Technologies, SACMAT '11, pp. 31–40. ACM, New York (2011)
2. American Psychiatric Association.: Committee on confidentiality. Guidelines on confidentiality. *Am. J. Psychiatry* **144**(11), 1522–1526 (1987)
3. Barnett, G., Barry, M., Robb-Nicholson, C., Morgan, M.: Overcoming information overload: an information system for the primary care physician. *Stud. Health Technol. Inform.* **107**(Pt 1), 273–276 (2004)
4. Beimel, D., Peleg, M.: Editorial: using OWL and SWRL to represent and reason with situation-based access control policies. *Data Knowl. Eng.* **70**(6), 596–615 (2011)
5. Biswas, A., Mynampati, K., Umashankar, S., Reuben, S., Parab, G., Rao, R., Kannan, V., Swarup, S.: MetDAT: a modular and workflow-based free online pipeline for mass spectrometry data processing, analysis and interpretation. *Bioinformatics* **26**(20), 2639–2640 (2010)
6. Blobel, B., Pharow, P.: Security and privacy issues of personal health. *Stud. Health Technol. Inform.* **127**, 288–297 (2007)
7. Blobel, B., Nordberg, R., Davis, J., Pharow, P.: Modelling privilege management and access control. *Int. J. Med. Inform.* **75**(8), 597–623 (2006)
8. Bouillon, Y., Wendling, F., Bartolomei, F.: Computer-supported collaborative work (CSCW) in biomedical signal visualization and processing. *IEEE Trans. Inform. Technol. Biomed.* **3**(1), 28–31 (1999)
9. Bradshaw, J., Uszok, A., Jeffers, R., Suri, N., Hayes, P., Burstein, M., Acquisti, A., Benyo, B., Breedy, M., Carvalho, M., Diller, D., Johnson, M., Kulkarni, S., Lott, J., Sierhuis, M., Hoof, R.V.: Representation and reasoning for DAML-based policy and domain services in KAoS and Nomads. In: Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS '03, pp. 835–842. ACM, New York (2003)
10. Bricon-Souf, N., Beuscart, R., Renard, J., Geib, J.: An asynchronous co-operative model for co-ordinating medical unit activities. *Comput. Methods Programs Biomed.* **54**(1–2), 77–83 (1997)
11. Buyl, R., Nyssen, M.: MedSkills: a learning environment for evidence-based medical skills. *Methods Inf. Med.* **49**(4), 390–395 (2010)
12. Calvillo, J., Roman, I., Roa, L.: Empowering citizens with access control mechanisms to their personal health resources. *Int. J. Med. Inform.* **82**(1), 58–72 (2013)
13. Chen, K., Chang, Y.C., Wang, D.W.: Aspect-oriented design and implementation of adaptable access control for electronic medical records. *Int. J. Med. Inf.* **79**(3), 181–203 (2010)
14. Croll, P.: Privacy, security and access with sensitive health information. *Stud. Health Technol. Inform.* **151**, 167–175 (2010)
15. Donelson, L., Tarczy-Hornoch, P., Mork, P., Dolan, C., Mitchell, J., Barrier, M., Mei, H.: The BioMediator system as a data integration tool to answer diverse biologic queries. *Stud. Health Technol. Inform.* **107**(Pt 2), 768–772 (2004)
16. D.T.C.S.E.C. (TCSEC).: DoD 5200.28-STD Foundations. MITRE Technical Report 2547 (1973)
17. Elkin, P., Liebow, M., Bauer, B., Chaliki, S., Wahner-Roedler, D., Bundrick, J., Lee, M., Brown, S., Froehling, D., Bailey, K., Famiglietti, K., Kim, R., Hoffer, E., Feldman, M., Barnett, O.: The introduction of a diagnostic decision support system (DXplain) into the workflow of a teaching hospital service can decrease the cost of service for diagnostically challenging diagnostic related groups (DRGs). *Int. J. Med. Inform.* **79**(11), 772–777 (2010)
18. Eriksson, H.: Using JessTab to integrate Protégé and Jess. *IEEE Intell. Syst.* **18**(2), 43–50 (2003)
19. Ferraiolo, D., Sandhu, R., Gavrila, S., Kuhn, D., Chandramouli, R.: Proposed NIST standard for role-based access control. *ACM Trans. Inf. Syst. Secur.* **4**(3), 224–274 (2001)

20. Ferreira, A., Correia, A., Silva, A., Corte, A., Pinto, A., Saavedra, A., Pereira, A., Pereira, A., Cruz-Correia, R., Antunes, L.: Why facilitate patient access to medical records. *Stud. Health Technol. Inform.* **127**, 77–90 (2007)
21. Ferrini, R., Bertino, E.: Supporting RBAC with XACML+OWL. In: Proceedings of the 14th ACM Symposium on Access Control Models and Technologies, SACMAT '09, pp. 145–154. ACM, New York (2009)
22. Finin, T., Joshi, A., Kagal, L., Niu, J., Sandhu, R., Winsborough, W., Thuraisingham, B.: ROWLBAC: representing role based access control in OWL. In: Proceedings of the 13th ACM Symposium on Access Control Models and Technologies, SACMAT '08, pp. 73–82. ACM, New York (2008)
23. Friedman, H.: Jess in Action: Java Rule-Based Systems. Manning Publications Co., Greenwich (2003)
24. Friedman, C., Wyatt, J.: Evaluation Methods in Biomedical Informatics, 2nd edn. Springer, New York (2006)
25. Geissbuhler, A.: Access to health information: a key for better health in the knowledge society. *Yearb. Med. Inform.* **2008**, 20–21 (2008)
26. Gennari, J., Weng, C., Benedetti, J., McDonald, D.: Asynchronous communication among clinical researchers: a study for systems design. *Int. J. Med. Inform.* **74**(10), 797–807 (2005)
27. Georgiadis, C., Mavridis, I., Pangalos, G., Thomas, R.: Flexible team-based access control using contexts. In: Proceedings of the 6th ACM Symposium on Access Control Models and Technologies, SACMAT '01, pp. 21–27. ACM, New York (2001)
28. Georgiadis, C., Mavridis, I., Nikolakopoulou, G., Pangalos, G.: Implementing context and team based access control in healthcare intranets. *Med. Inform. Internet Med.* **27**(3), 185–201 (2002)
29. Gouglidis, A., Mavridis, I.: domRBAC: an access control model for modern collaborative systems. *Comput. Secur.* **31**(4), 540–556 (2012)
30. Grando, M., Peleg, M., Cuggia, M., Glasspool, D.: Patterns for collaborative work in health care teams. *Artif. Intell. Med.* **53**(3), 139–160 (2011)
31. Halsted, M., Perry, L., Cripe, T., Collins, M., Jakobovits, R., Benton, C., Halsted, D.: Improving patient care: the use of a digital teaching file to enhance clinicians' access to the intellectual capital of interdepartmental conferences. *AJR. Am. J. Roentgenol.* **182**(2), 307–309 (2004)
32. Hannan, A.: Providing patients online access to their primary care computerised medical records: a case study of sharing and caring. *Inform. Prim. Care* **18**(1), 41–49 (2010)
33. Hoelzer, S., Schweiger, R., Rieger, J., Meyer, M.: Dealing with an information overload of health science data: structured utilisation of libraries, distributed knowledge in databases and Web content. *Stud. Health Technol. Inform.* **124**, 549–554 (2006)
34. Hu, V., Scarfone, K.: Guidelines for access control system evaluation metrics. National Institute of Standards and Technology. Interagency Report 7874 (2012)
35. Hu, V., Ferraiolo, D., Kuhn, D.: Assessment of access control systems. National Institute of Standards and Technology. Interagency Report 7316 (2006)
36. IBM SPSS Software.: <http://www.ibm.com/software/analytics/spss/>. Accessed 1 Aug 2014
37. Jess, the Rule Engine for the JavaTM Platform. <http://www.jessrules.com>. Accessed 1 Aug 2014
38. Jitaru, E., Moisil, I., Alexandru, A., Mirescu, M., Pertache, I.: CSCW – a paradigm for an efficient management of the healthcare organizations. *Stud. Health Technol. Inform.* **90**, 596–600 (2002)
39. Jung, Y., Joshi, J.: CPBAC: property-based access control model for secure cooperation in online social networks. *Comput. Secur.* **41**, 19–39 (2014)
40. Kamateri, E., Kalampokis, E., Tambouris, E., Tarabanis, K.: The linked medical data access control framework. *J. Biomed. Inform.* **50**, 213–225 (2014)
41. Kesselheim, A., Mello, M.: Confidentiality laws and secrecy in medical research: improving public access to data on drug safety. *Health Aff.* **26**(2), 483–491 (2007)
42. Kopena, K., Regli, W.: DAMLJessKB: a tool for reasoning with the semantic web. *IEEE Intell. Syst.* **18**(3), 74–77 (2003)
43. Kopsacheilis, E., Kamilatos, I., Strintzis, M., Makris, L.: Design of CSCW applications for medical teleconsultation and remote diagnosis support. *Med. Inform.* **22**(2), 121–132 (1997)

44. Koufi, V., Vassilacopoulos, G.: Context-aware access control for pervasive access to process-based healthcare systems. *Stud. Health Technol. Inform.* **136**, 679–684 (2008)
45. Kunzi, J., Koster, P., Petkovic, M.: Emergency access to protected health records. *Stud. Health Technol. Inform.* **150**, 705–709 (2009)
46. Kurtz, G.: EMR confidentiality and information security. *J. Healthc. Inf. Manag.* **17**(3), 41–48 (2003)
47. Lampson, B.: Dynamic protection structures. In: Proceedings of the November 18–20, 1969, Fall Joint Computer Conference, AFIPS '69 (Fall), pp. 27–38. ACM, New York (1969)
48. LaPadula, L., Bell, D.: Secure Computer Systems: Mathematical Foundation, vol. 1. Hansom AFB, Bedford (1973)
49. Le, X., Wang, D.: Development of a system framework for implementation of an enhanced role-based access control model to support collaborative processes. In: Proceedings of the 3rd USENIX Conference on Health Security and Privacy, HealthSec'12, pp. 9–9 (2012)
50. Le, X., Lee, S., Lee, Y.K., Lee, H., Khalid, M., Sankar, R.: Activity-oriented access control to ubiquitous hospital information and services. *Inform. Sci.* **180**(16), 2979–2990 (2010)
51. Le, X., Doll, T., Barbosu, M., Luque, A., Wang, D.: An enhancement of the role-based access control model to facilitate information access management in context of team collaboration and workflow. *J. Biomed. Inform.* **45**(6), 1084–1107 (2012)
52. Le, X., Doll, T., Barbosu, M., Luque, A., Wang, D.: Evaluation of an enhanced role-based access control model to manage information access in collaborative processes for a statewide clinical education program. *J. Biomed. Inform.* **50**, 184–195 (2014)
53. Lee, E., McDonald, D., Anderson, N., Tarczy-Hornoch, P.: Incorporating collaborative concepts into informatics in support of translational interdisciplinary biomedical research. *Int. J. Med. Inform.* **78**(1), 10–21 (2009)
54. Lewis, D.: Information overload: tips for focusing on what you need and ignoring what you don't. *Biomed. Instrum. Technol.* **43**(3), 188–195 (2009)
55. Lindberg, D., Humphreys, B.: Rising expectations: access to biomedical information. *Yearb. Med. Inform.* **2008**, 165–172 (2008)
56. Lovis, C., Spahni, S., Cassoni, N., Geissbuhler, A.: Comprehensive management of the access to the electronic patient record: towards trans-institutional networks. *Int. J. Med. Inform.* **76**(5–6), 466–470 (2007)
57. Maas, J., Kamm, W., Hauck, G.: An integrated early formulation strategy – from hit evaluation to preclinical candidate profiling. *Eur. J. Pharm. Biopharm.* **66**(1), 1–10 (2007)
58. Mikels, D.: Privacy: after the compliance date. *J. Healthc. Inf. Manag.* **18**, 34–37 (2007)
59. Moraitis, P., Petraki, P., Spanoudakis, N.: Engineering jade agents with the Gaia methodology. In: Agent Technologies, Infrastructures, Tools, and Applications for E-Services, vol. 2592, pp. 77–91. Springer, Heidelberg (2002)
60. Motta, G., Furui, S.: A contextual role-based access control authorization model for electronic patient record. *IEEE Trans. Inform. Technol. Biomed.* **7**(3), 202–207 (2003)
61. New York State HIV Clinical Education Initiative. <http://www.ceitraining.org>. Accessed 1 Aug 2014
62. Ni, Q., Bertino, E., Lobo, J., Brodie, C., Karat, C.M., Karat, J., Trombeta, A.: Privacy-aware role-based access control. *ACM Trans. Inf. Syst. Secur.* **13**(3), 24:1–24:31 (2010)
63. Niazkhani, Z., Pirnejad, H., van der Sijs, H., de Bont, A., Aarts, J.: Computerized provider order entry system—does it support the inter-professional medication process? Lessons from a Dutch academic hospital. *Methods Inform. Med.* **49**(1), 20–27 (2010)
64. Nijhuis, B., Reinders-Messelink, H., de Blecourt, A., Olijve, W., Groothoff, J., Nakken, H., Postema, K.: A review of salient elements defining team collaboration in paediatric rehabilitation. *Clin. Rehabil.* **21**(3), 195–211 (2007)
65. OASIS: eXtensive Access Control Markup Language (XACML) Version 3.0. <http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-cd-1-en.html>. Accessed 1 Aug 2014
66. O'Connor, M., Knublauch, H., Tu, S., Grososf, B., Dean, M., Grossos, W., Musen, M.: Supporting rule system interoperability on the semantic web with SWRL. In: Proceedings of the 4th International Conference on The Semantic Web, ISWC'05, pp. 974–986. Springer, Heidelberg (2005)

67. Park, J., Sandhu, R.: The UCONABC usage control model. *ACM Trans. Inf. Syst. Secur.* **7**(1), 128–174 (2004)
68. Peleg, M., Beimel, D., Dori, D., Denekamp, Y.: Situation-based access control: privacy management via modeling of patient data access scenarios. *J. Biomed. Inform.* **41**(6), 1028–1040 (2008)
69. Pratt, W., Reddy, M., McDonald, D., Tarczy-Hornoch, P., Gennari, J.: Incorporating ideas from computer-supported cooperative work. *J. Biomed. Inform.* **37**(2), 128–137 (2004)
70. Predeschlyl, M., Dadam, P., Acker, H.: Security challenges in adaptive e-Health processes. In: Proceedings of the 27th International Conference on Computer Safety, Reliability, and Security, SAFECOMP '08, pp. 181–192. Springer, Heidelberg (2008)
71. Protégé. <http://protege.stanford.edu>. Accessed 1 Aug 2014
72. Renegar, G., Webster, C., Stuerzebecher, S., Harty, L., Ide, S., Balkite, B., Rogalski-Salter, T., Cohen, N., Spear, B., Barnes, D., Brazell, C.: Returning genetic research results to individuals: points-to-consider. *Bioethics* **20**(1), 24–36 (2006)
73. Reynolds, R., Candler, C.: MedEdPORTAL: educational scholarship for teaching. *J. Contin. Educ. Health Prof.* **28**(2), 91–94 (2008)
74. Rinehart-Thompson, L., Hjort, B., Cassidy, B.: Redefining the health information management privacy and security role. *Perspect. Health Inf. Manag.* **6**, 1–11 (2009)
75. Rodriguez, M., Favela, J., Martinez, E., Munoz, M.: Location-aware access to hospital information and services. *IEEE Trans. Inf. Technol. Biomed.* **8**(4), 448–455 (2004)
76. Ross, S., Lin, C.T.: The effects of promoting patient access to medical records: a review. *J. Am. Med. Inform. Assoc.* **10**(2), 129–138 (2003)
77. Sittig, D., Singh, H.: Eight rights of safe electronic health record use. *J. Am. Med. Assoc.* **302**(10), 1111–1113 (2009)
78. Sousa, A., Wagner, D., Henry, R., Mavis, B.: Better data for teachers, better data for learners, better patient care: college-wide assessment at Michigan State University's College of Human Medicine. *Med. Educ. Online* **16**, 1–10 (2011)
79. Sujansky, W., Faus, S., Stone, E., Brennan, P.: A method to implement fine-grained access control for personal health records through standard relational database queries. *J. Biomed. Inform.* **43**(5 Suppl), S46–S50 (2010)
80. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. <http://www.w3.org/Submission/SWRL/>. Accessed 1 Aug 2014
81. The Workflow Engine Model. [https://msdn.microsoft.com/en-us/library/aa188337\(office.10\).aspx](https://msdn.microsoft.com/en-us/library/aa188337(office.10).aspx). Accessed 1 Aug 2014
82. Toninelli, A., Montanari, R., Kagal, L., Lassila, O.: A semantic context-aware access control framework for secure collaborations in pervasive computing environments. In: Proceedings of the 5th International Conference on The Semantic Web (ISWC), ISWC'06, pp. 473–486. Springer, Heidelberg (2006)
83. Unertl, K., Weinger, M., Johnson, K., Lorenzi, N.: Describing and modeling workflow and information flow in chronic disease care. *J. Am. Med. Inform. Assoc.* **16**(6), 826–836 (2009)
84. Van Harrison, R., Standiford, C.J., Green, L.A., Bernstein, S.J.: Integrating education into primary care quality and cost improvement at an academic medical center. *J. Cont. Educ. Health Prof.* **26**(4), 268–284 (2006)
85. Wang, E., Kim, Y.: A teaching strategies engine using translation from SWRL to Jess. In: Proceedings of the 8th International Conference on Intelligent Tutoring Systems (ITS), pp. 51–60 (2006)
86. Wang, D., Peleg, M., Bu, D., Cantor, M., Landesberg, G., Lunenfeld, E., Tu, S., Kaiser, G., Hripcsak, G., Patel, V., Shortliffe, E.: GESDOR – a generic execution model for sharing of computer-interpretable clinical practice guidelines. In: AMIA Annual Symposium Proceedings, pp. 694–698 (2003)
87. Wang, D., Peleg, M., Tu, S., Boxwala, A., Ogunyemi, O., Zeng, Q., Greenes, R., Patel, V., Shortliffe, E.: Design and implementation of the GLIF3 guideline execution engine. *J. Biomed. Inform.* **37**(5), 305–318 (2004)

88. Xiao, Y., Seagull, F.: Emergent CSCW systems: the resolution and bandwidth of workplaces. *Int. J. Med. Inform.* **76**(Suppl 1), S261–S266 (2007)
89. Yeh, L.Y., Chen, Y.C., Huang, J.L.: ABACS: an attribute-based access control system for emergency services over vehicular ad hoc networks. *IEEE J. Sel. Areas Commun.* **29**(3), 630–643 (2011)
90. Zhang, L., Ahn, G.J., Chu, B.T.: A rule-based framework for role-based delegation and revocation. *ACM Trans. Inf. Syst. Secur.* **6**(3), 404–441 (2003)
91. Zhang, X., Oh, S., Sandhu, R.: PBDM: a flexible delegation model in RBAC. In: Proceedings of the 8th ACM Symposium on Access Control Models and Technologies, SACMAT '03, pp. 149–157. ACM, New York (2003)

Chapter 14

Automating Consent Management Lifecycle for Electronic Healthcare Systems

Muhammad Rizwan Asghar and Giovanni Russello

Abstract The notion of patient's consent plays a major role in granting access to medical data. In typical healthcare systems, consent is captured by a form that the patient has to fill-in and sign. In e-Health systems, the paper-form consent is being replaced by access control mechanisms that regulate access to medical data, while taking into account electronic content. This helps in empowering the patient with the capability of granting and revoking consent in a more effective manner. However, the process of granting and revoking consent greatly varies according to the situation in which the patient is. Our main argument is that such a level of detail is very difficult and error-prone to capture as a set of authorisation policies. In this chapter, we present ACTORS (Automatic Creation and lifecycle managemenT Of authoRisation policieS), a goal-driven approach to manage consent. The main idea behind ACTORS is to leverage the goal-driven approach of Teleo-Reactive (TR) programming for managing consent that takes into account changes regarding the domains and contexts in which the patient is providing her consent.

14.1 Introduction

Healthcare information refers to any data containing information about an individual's health conditions. As it contains sensitive personal information, its improper disclosure may influence several aspects of an individual's life. Today, medical data is massively being converted into electronic format. Individuals' medical data can be now easily accessible to a very large number of health-care professionals. Although this is done with the best of intentions to improve the processing and streamline healthcare delivery, it also poses very concrete threats to the individual's privacy.

This chapter extends our work that appeared in the Proceedings of POLICY 2012 [3].

M.R. Asghar (✉) • G. Russello

Department of Computer Science, The University of Auckland, Auckland, New Zealand

e-mail: r.asghar@auckland.ac.nz; g.russello@auckland.ac.nz

Since the medical information of an individual is confidential, the only basis for accessing it is through that individual's consent. In traditional healthcare systems, an individual provided her consent by signing a paper form. In these settings, withdrawing consent was very difficult for an individual because she had to go through complicated bureaucratic processes. Moreover, the granularity of consent was very coarse-grained. The individual agreed in providing consent in advance for all her medical data, thus violating the principle of least privilege [36]—a principle that advocates for providing only legitimate access to requested resources for a limited time necessary to complete the job.

Policy-based authorisation mechanisms have successfully been used in managing access rights given the flexibility and re-usability that they offer. In literature, several approaches have been realised where the notion of consent is integrated with the policy decision mechanism. For instance, Russello et al. [35] propose to capture the notion of consent through the use of medical workflow and to integrate it with Ponder2 authorisation policies [41]. Ponder2 authorisation policies are represented as a (S, A, T) tuple, meaning a subject S can take action A on target T. For instance, a nurse (S) can read (A) patients' records (T). Wuyts et al. [43] have extended the eXtensible Access Control Markup Language (XACML) [30] authorisation model with the notion of consent. XACML is an eXtensible Markup Language (XML)-based language designed for specifying fine-grained access control. It is a standard ratified by the Organisation for the Advancement of Structured Information Standards (OASIS). XACML policies are expressed as a set of rules for regulating access to the resources. A XACML request, containing necessary information in making authorisation decision, is evaluated against XACML policies.

To specify a set of authorisation policies that capture all the details required to enforce correctly an individual's decisions about consent is very complex. First of all, each authorisation policy has conditions to express when it should be enforced that might be in conflict with other policies. Although work has been done to address the problem of automatically resolving conflicts [34], it is not possible to completely automate the decision since in the specific case of the healthcare scenario, humans are also involved. To complicate matters further, contextual information needs to be captured to identify the purpose of the access being requested. If these details are not captured correctly in the policy specification by the security administrator then there may be serious consequences.

For instance, the way in which an individual wants to provide and revoke her consent differs according to the caregivers that she is interacting with. With her General Practitioner (GP), a patient typically establishes a lasting relationship; therefore, consent can be given for a long time. On the other hand, when she is visiting a specialist in a hospital, she wants to give consent only for the time the treatment will last and only for the data that is required for the specific treatment. Still, another different situation is in the case of an emergency where the paramedics have to provide first care before reaching the emergency room. In this case, consent can be given to the medical data however for the short period of time required to reach the hospital.

From the above scenario, it emerges that specifying in one single policy set all the requirements for managing consent is a very error-prone task. Moreover, as argued in [39] users are not engaging with their privacy tools and often prefer to ignore them. In the light of this, in this chapter we propose ACTORS (Automatic Creation and lifecycle managemenT Of authoRisation policieS), where a goal-driven approach is used to *glue* together and manage authorisation policies that have a common aim, that is the handling of consent in a specific context (i.e., consent for the GP, for the specialist, and paramedics). In particular, our observation is that we can simplify the specification of authorisation policies when these are treated as a *program sequence* towards a specific goal. The main contribution and novelty of our approach is to propose the idea of using Teleo-Reactive (TR) programs to glue together authorisation policies aiming at a specific goal. The idea of TR programs was introduced by Nilsson [28]. The main advantage of TR programs is that the way in which they are specified is very natural for humans. Therefore, a security administrator can capture more naturally the security requirements in a TR sequence.

The rest of this chapter is organised as follows. In Sect. 14.2, we discuss the legal aspect related to consent and set some of the terminology that will be used in the rest of this chapter. Section 14.3 describes an overview of a case study that we use to demonstrate the feasibility of our approach. Next, we provide a brief overview of TR Policies in Sect. 14.4. In Sect. 14.5, we present our proposed approach. In Sect. 14.6, we show how the case study scenarios can be modelled using the proposed approach. Related approaches are reviewed in Sect. 14.7. Finally, we conclude and indicate some directions for future work in Sect. 14.8.

14.2 Legal Background

In this section, we will discuss some of legal frameworks related to data privacy and consent. This is not intended to be an exhaustive discussion on all the legal frameworks out there. On the other hand, we feel it is necessary to put within the law perspective the technical discussion that will follow in this chapter.

14.2.1 Legal Framework for Consent

When dealing with people's data, the most developed countries have established legal frameworks to provide individuals with rights to allow them to make decisions regarding collection, use and disclosure of personal data. As discussed in [39], this approach can be considered as a "privacy self-management" and relies entirely on the user to take decisions and actions to either protect or disclose her data.

In the last 25 years, to deal with the growing demand and new capabilities for data collection and aggregation, a significant number of new laws have been proposed and passed in the U.S.; these include Organisation for Economic Co-operation and Development (OECD) Privacy Guidelines in 1980, the Asia-Pacific Economic Cooperation (APEC) Privacy Framework in 2004, and more recently in 2012 the Federal Trade Commission (FTC) and the White House issued major new frameworks for protecting privacy. All these efforts have in common a set of principles for protecting privacy that was first proposed in the Fair Information Practice Principles (FIPPs) as a report by the U.S. Department of Health, Education, and Welfare in 1973 [33]. The FIPPs include several guidelines such as (1) transparency of the record system of personal data; (2) the right to notice; (3) the right to prevent the use of personal data for new purposes without consent; (4) the right to correct/amend one's record; and (5) responsibility of the data holder for protecting data from misuse.

The privacy law framework in Canada is also based on the OECD Guidelines proposed in 1980 and relies on consent for collection, use and disclosure of personal data. However, as discussed in [20], Canada's legislation on data handling and processing makes a clear separation on role of consent between public and private sectors. In particular, in the public sector, consent is seen as a *justification* whether in the private sector consent is a *requirement* for collecting, using and disclosing data. The Australian Privacy Act provides general guidelines for collecting and processing personal data that are also based on consent. However, for larger commercial entities (with annual income greater than three million AUD), there is also an extra burden to destroy personal data once it has been used as intended at collection time. For instance, if a customer's address was collected because of the delivery of goods, once the goods are delivered then the information should be destroyed. New Zealand has a more relaxed approach when it comes to collection of data for commercial purpose. For instance, an entity can collect information about an individual as long as the individual has been informed about the collection and the purpose for which the data has been collected. More interestingly, an entity does not need to inform again an individual if the same type of data is collected again after the person has been correctly informed the first time. However, in New Zealand, user consent for collecting, using, and disclosing personal data is required in specific sectors related to healthcare, telecommunications and credit records.

Compared to the frameworks of the countries above, the EU directives have a more paternalistic approach when it comes to data processing, as discussed in [37]. According to article 2(h) of the EU Data Protection Direction (DPD) [12], consent is defined as: "*the data subject's consent shall mean any freely given specific and informed indication of his wishes by which a data subject signifies his agreement to personal data relating to him being processed*", where the term *data subject* describes an individual whose data is handled and *data controller* indicates any entity that handles personal data.

The EU law framework also supports the concept of privacy self-management. According to article 7 (a) of the EU Data Protection Direction [12], a data subject's personal data may only be processed if she has given her consent. However,

the way in which a data handler seeks the data subject's consent is much more regulated. Furthermore, data processing in the EU is always controlled through a legal framework; whereas, in the U.S., data processing is always granted by default unless explicitly forbidden by the law.

14.2.2 Consent in Healthcare Systems

In the context of healthcare systems, consent indicates agreement of the patient on sharing her personal health information [32]. In traditional healthcare systems, a data subject provides her paper-based consent typically once she is enrolled within the system. Generally, the paper-based consent is considered valid once signed by the data subject. Unfortunately, there are two main problems with the paper-based consent. First, it becomes very cumbersome for the data subject to withdraw her paper-based consent. That is, she has to go through complicated bureaucratic processes where she has to call on the responsible authority to withdraw her consent with some considerable effort, waste of time and a huge sense of frustration. Second, a data subject provides her consent in advance for all her medical data at the time of registration with the healthcare system even when it may not be necessarily used, thus violating the principle of least privilege.

With the introduction of electronic healthcare systems, we have moved from the paper-based consent to the *electronic consent*, or *e-consent* in short. e-consent has been established as a new industry standard [26] and aims at replacing the traditional paper-based control, thus providing more control to patients for controlling the way they share their Electronic Health Records (EHR).

In current IT healthcare systems, the notion of e-consent is captured as *authorisation policies* that control the access to the data, such as in [35]. Technically, the creation or editing of these authorisation policies is delegated to an IT security administrator. The security administrator operates on behalf of the data subject to deploy policies in the IT infrastructure of the data controller. In some countries, specific legislation may require the digital consent to be digitally signed by the data subject to be considered equivalent to the manually signed paper-based consent [7].

Using the classification proposed in [5], it is possible to identify the following essential elements of e-consent:

Requester: An entity to whom the authorisation is provided. It could be a person, a role or even an organisation.

Actions: These are the set of rights that are authorised by the consent.

Purpose: A purpose is a reason for which the authorisation is given.

Validity: It is a time period in which the authorisation is applicable.

Revocation: This is a feature of consent using which one can revoke her consent.

Delegation: Delegation is an authority given to someone who can manage consent on behalf of someone else.

In the following discussion, we will use the term consent to indicate in general electronic consent. We can identify two categories of consent [6, 32]: implicit and explicit consent. Implicit (a.k.a. implied) consent is one that is inferred from the actions. Explicit consent is one that is given explicitly. There is another type of consent called informed consent. In the context of healthcare systems, informed consent requires patients to be informed about what they are going to agree with. In [6], Coiera and Clarke list the following four forms of consent:

1. **General Consent:** This is one time consent given by a patient to any medical professional, for any purpose and valid as long as it is not revoked by the patient. This form ensures ease of use but might hamper protection due to open access.
2. **General Consent with Specific Denials:** The patient provides a general consent except some specific conditions that could be based on expressive policies e.g., based on time, purpose and/or validity. From the security point of view, it improves the general consent form but reduces availability.
3. **General Denial with Specific Consent(s):** It is opposite of the previous form. In this case, the patient provides a general denial except specific conditions under which she would like to give her consent. This consent type provides reasonable control as well as restricted availability.
4. **General Denial:** It is opposite of the first form. In this case, a patient denies access to her personal information.

There are two major factors affecting the evaluation criteria for the consent. First, it is the *ease of use*, meaning how easy it is to access and use the consent. The second one is *privacy*, which means the level of protection offered by consent. If we evaluate the above four forms of consent under this evaluation criteria then the general denial provides better protection of privacy and this level decreases as we choose other forms (from bottom to up) of consent. Thus, the protection of privacy provided by the general consent is at the lowest level among all other forms. However, the general consent ensures the ease of access. The ease of access decreases as we move down from the general consent to the general denial.

It is important to know that consent is required for providing access to medical data that is not anonymised. However, consent might not be required when the patients' data is first anonymised and then shared, given the data anonymisation technique can guarantee privacy of the patients.

14.2.3 Consent Limitations

In practice, these law frameworks rely on the data subjects to make decisions on whether it is beneficial to them to consent access and usage of their data: consent legitimises any collection, use and disclosure of someone's data.

There are several limitations with this approach. First of all, there are cognitive problems with privacy self-management. As shown by several empirical studies in social science research, people do not engage with privacy self-management: the

main reason is that they do not read the terms and conditions notices [29] or when they read them they do not understand them [11]. However, even when people read and understand the notices they are not able to take rational and informed decisions on the costs and benefits on consenting access to their data [40].

Another issue is more related to the scale of the problem. Assuming that people had a complete understanding of the risks involved in consenting access to their personal data, there are far too many entities collecting data to make self-managing privacy practically possible. One study has estimated that the cost associated with the lost productivity if each of us were to read the terms and conditions notices of each website we visit on a given year to a staggering \$781 billion [25]. To complicate matters even further, each entity quite frequently changes privacy policies, which would require further engagement from the user.

Third, the major harm to one's privacy comes from the aggregation of data collected by different parties over a period of time. It is almost impossible for a user to be able to understand the risks and benefits at the time of the release of a piece of information without a proper knowledge of how the data will be used in aggregation with other information. For instance, several entities that have received consent to use the data subject's data, that in isolation and at a given point in time are not harmful, could decide later to collaborate or aggregate their data resulting in a violation of the data subject's privacy. How can a data subject be able to predict such an event at the time the consent is given? Privacy regulations aim mainly at protecting one's Personally Identifiable Information (PII). However, PII is not a static label that can be associate to a piece of data for its entire life cycle. With the huge amount of facts that we leave in our digital trail online and the advancements of data mining technologies, identifying someone is becoming very easy from data that taken in isolation is pretty harmless [44]. The result is that with data mining and aggregation, it is nearly impossible to be able to manage one's personal information.

Another negative aspect of privacy self-management is that it is always considered as an isolated transaction between an individual and an entity. However, some aspects of the privacy have an effect not only on the individual but to a society as a whole. The decisions taken by an individual for consenting collection, usage and disclosure of her data might not have the most desirable effect on a larger scale. On the other hand, sometime overriding one's privacy can be beneficial for the protection or advantages of our society. For instance, as discussed in [38], the use of data analytics can lead to better medical treatments as well as better responses to data breaches. Privacy self-management fails to address the global outcomes on a social level, focusing only on the single individual and on isolated transactions. Last but not least, data subjects may withdraw their consent at any time.

Considering the criticisms above, one would be tempted to abandon consent as a mean to safeguard one's privacy. However, we argue that controlling consent could be improved by bringing the necessary tools to access and manage consent closer to the data subject, such as to her mobile device. Also, instead of doing consent micro management by presenting endless requests of binary consent decisions, we want to provide the data subject with an approach that takes into account her goals when it comes to protect her privacy, and that will learn from the decisions the data subject takes in a particular context.

14.3 A Case Study

In this section, we introduce the case study that we will use throughout the paper to demonstrate the feasibility of our approach. The case study is partially inspired from the European funded projects including ENDORSE [23] and EnCoRe [42]. Both projects focus on developing IT solutions for privacy preserving data management, where *consent* is one of the main point of focus.

In this section, we describe several scenarios based on the IT healthcare system currently deployed in one of the major hospitals in Italy. We assume that each patient has a smartphone that she uses to receive requests for giving her consent when she is interacting with the medical personnel. A patient can review through her smartphone who is requesting the access, the purpose of the request, and which data is requested.

At the time of providing consent, a patient may decide to save her preferences for subsequent consent requests made in the same context and/or by the same entity. Afterwards, a patient may withdraw her saved preferences regarding consent. Furthermore, a patient may activate withdrawn preferences regarding her consent. Last but not least, a patient may intend to delete, forever, her preferences, initially saved for providing consent automatically.

Patient Visiting Her GP Let us consider the healthcare scenario where Alice moves to Milan and visits her GP for the first time. The GP requires access to Alice's medical history consisting of several medical tests and reports. For this purpose, the GP requires Alice's consent. Alice receives the consent request on her smartphone and decides to provide her consent also in the future.

Patient Visiting a Cardiologist Later, the GP of Alice discovers that she has a heart disorder. In this case, the GP refers Alice to a cardiologist for further testing. For visiting the cardiologist, Alice needs to contact the hospital booking service for getting an appointment. The hospital has several cardiologists; thus, it is not known in advance which one is assigned prior to the actual appointment. On the day of appointment, Alice will know the assigned cardiologist and can consent the cardiologist to access her medical data. However, Alice's consent should be valid for the duration of the treatment and the data accessed should be within the scope of the treatment (i.e., the cardiologist should not have access to Alice's gynaecological reports). Moreover, if Alice is not happy with the assigned cardiologist then she may withdraw her consent and request a new cardiologist.

Patient in an Emergency Situation While Alice is driving in her car, she has a car accident and gets injured. The emergency response team reaches the accident location and starts treating Alice. For the treatment, the paramedic requires Alice's consent to access her medical history to get information about her allergies and any serious conditions that she already may have. Alice provides consent to access her medical records so that the paramedic is aware of her heart problem and provides the appropriate treatment that does not interfere with the treatment prescribed by the cardiologist. Although the paramedic has access to Alice's full medical record, consent should be revoked when the emergency is over.

14.4 Overview of Teleo-Reactive Policies

From the above scenarios, it is clear that to capture all the details required to express the data subject's consent in different settings is very complex. If these details are not captured correctly by the security administrator in the policy specification then serious consequences might happen. In our experience, capturing all the security requirements through the specification of several independent authorisation policies is a very hard task. In the specific case of capturing a data subject's consent, it becomes even more complicated since there is the involvement of a human (which is the data subject that can grant, hold and withdraw consent) and contextual information expressed in the policies (such as the location and time of the access).

In this chapter, we propose to employ a goal-driven approach to *glue* together and manage authorisation policies that have a common aim, i.e., the handling of consent. In particular, our observation is that we can simplify the specification of authorisation policies when these are treated as a *program sequence* towards a specific goal. In this chapter, we propose to leverage the idea of TR programs to glue together authorisation policies aiming at a specific goal. The idea of TR programs was initially introduced by Nilsson [28]. A TR program is a control sequence directing towards a goal while taking into account changes in environmental circumstances. TR programs were used for automating behavioural robotics where a robot was continuously observing its environmental changes.

In the following, we provide a brief overview of TR policies that is similar to one introduced by Marinovic et al. in [24].

14.4.1 TR Policy Representation

A TR policy is an ordered list of rules as shown in Fig. 14.1, where each rule contains (Line 2) a condition part and an action part. The condition part contains a predicate that is bound with a variable, which is denoted with V . These variables may describe facts or states of the system or environment in which a TR policy is evaluated. A variable starts with a capital letter while a condition or an action starts with a small letter. The action part contains a function that is called by the TR policy. The action part may contain variables. The condition and action parts are separated

```

1 tr policy name( $P_1, P_2, \dots, P_m$ )
2 cond1( $V$ ) → action1( $V$ )
3 cond2_a ∧ (cond2_b ∨ ¬cond2_c) → action2_y ⊗ action2_z
4 cond3( $P_1$ ) → action3_a || action3_b
5 ...
6 condn_1 ∧ condn_2 ... ∨ condn_x → actionn_1 || actionn_2 ... ⊗ actionn_y
```

Fig. 14.1 A layout of TR policies

```

1 tr policy superStore(E)
2
3  isStoreCrowded ∧ isAvailable(CC) → serverAtCheckoutCounter(E,CC)
4
5  askedForHelp(E,C) → helpCustomer(E,C)
6
7  isEmpty(S) → stackShelf(E,S)

```

Fig. 14.2 An example of a TR policy

by \rightarrow . Each TR-policy has a name starting with a small letter and can be instantiated with some parameters (Line 1). The condition part may include parameters, each denoted by P_i (Line 4). The condition part can contain either a single condition or form (Line 2 or Line 6) a conditional expression where multiple conditions can be combined using logical operators \wedge and \vee . Similarly, the action part can contain either a single function or multiple functions that may be executed sequentially and/or concurrently. The sequential and concurrent execution of functions can be represented with \otimes operator (Line 3 and Line 6) and \parallel operator (Line 4 and Line 6), respectively. In a TR policy, rules are specified in the descending order with respect to their priorities. That is, a high priority rule comes first.

In Fig. 14.2, we have illustrated an example of a TR policy representing job specification of an employee E who works at a superstore. For simplicity, we mainly consider three job responsibilities: serving at checkout counters, helping customers and stacking shelves. The top priority will be given to serving at a checkout counter. An employee E can serve at a checkout counter CC when each occupied checkout counter is crowded by a large number of customers. Of course, a checkout counter CC should be available before an employee E can start serving. This job responsibility is specified as the first rule at Line 3. The next priority will be given to helping any customer C, meaning if an employee E is asked for any help then she should help the customer C. The second rule (Line 5) in the TR policy represents this job responsibility. The lowest priority job responsibility is stacking a shelf S if it is empty (or about to empty)—see the last rule (Line 7) in the TR policy.

14.4.2 TR Policy Evaluation

The runtime of the TR policy monitors changes in facts or states about the system or environment in which evaluation is performed. These changes can result in the condition part of a rule becoming either *true* or *false*. The functions in the action part of a rule will be executed if its condition part is evaluated to *true* by the runtime. In a TR policy, the condition part corresponding to the highest priority rule is evaluated first. If it evaluates to *false*, the condition part of the next high priority rule will be evaluated. In other words, if the action part of any rule is being executed, it means the condition parts of all higher priority rules (as compared to the current rule) are

evaluated to *false*. The action part of any rule is executed as long as its condition part evaluates to *true* while condition parts of all higher priority rules (as compared to the current rule) remain *false*.

As an example, let's discuss evaluation of the TR policy illustrated in Fig. 14.2. This TR policy consists of three rules listed in the order of priority (from higher to lower). The runtime of the TR policy checks if the superstore is crowded, i.e., *isStoreCrowded*. Then, it identifies whether any checkout counter is available, i.e., *isAvailable(CC)*. If both conditions are met (i.e., they evaluate to *true*), the first rule is fired and the employee E starts serving, i.e., *serverAtCheckoutCounter(E, CC)*. However, if the superstore is not crowded (i.e., *isStoreCrowded* evaluates to *false*) or it is crowded but no checkout counter is available (i.e., *isAvailable(CC)* evaluates to *false*) then the runtime will start evaluating conditions of next rules in the TR policy. In our case, the next rule is to monitor if the employee E is asked for help by any customer C (i.e., *askedForHelp(E, C)*). If so (i.e., *askedForHelp(E, C)* evaluates to *true*), the employee will start serving the customer (i.e., *helpCustomer(E, C)*). Otherwise, the runtime will evaluate condition of the last rule, i.e., if a shelf is empty. It will evaluate *isShelfEmpty(S)*. If it is *true*, the employee E will start stacking the shelf S (i.e., *stackShelf(E, S)*). As soon as execution of any rule is completed, the runtime will start evaluating condition of the first rule in the TR policy (i.e., serving at a checkout counter if the superstore is crowded).

14.5 The ACTORS Approach

ACTORS aims at automating creation and management of consent related authorisation policies using a goal-driven approach. Figure 14.3 illustrates the ACTORS architecture. There are three main system entities:

Data Subject: Data subjects represent end-users and are key entities, responsible for granting, updating or withdrawing their consent.

Data Requester: A data requester is an entity who makes a consent request. In the context of healthcare systems, this entity could be a doctor or a GP.

Smartphone: It is a mobile device that is owned, or at least operated, by data subjects. It automatically manages lifecycle of consent authorisation policies.

In Fig. 14.3, an initial consent request is issued by the data subject as we can see in Step (i). The Administration Point, managed by the data subject's smartphone, receives the request and fetches corresponding TR and template policies from the TR and template store in Step (ii). The main idea is that each TR policy captures a specific goal, such as managing consent for the GP. TR policies are used for instantiating authorisation policies from a standard set of policy templates. TR policies also manage the lifecycle of instantiated authorisation policies.

Since all the details required in an authorisation policy may not be known in advance (such as, ID of the specific cardiologist assigned on the day of the visit, location where the visit will take place), we use policy templates to define

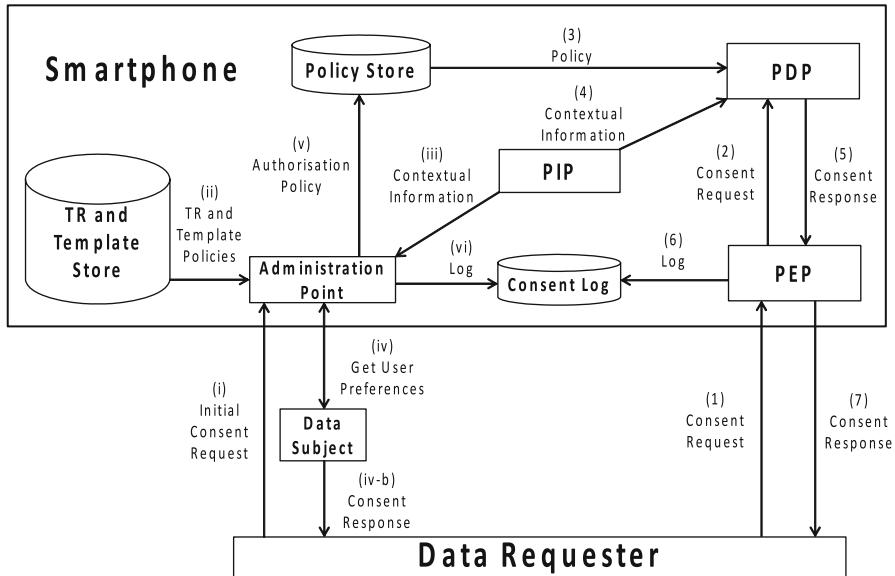


Fig. 14.3 The ACTORS architecture for managing consent lifecycle

abstract authorisation policies. When all the required information is available, TR policies can instantiate the required authorisation policies from the given templates. This instantiated authorisation policy is stored and enforced by the data subject's smartphone, thus providing greater control to data subjects to manage their consent.

Authorisation policies are created and managed based on the data subject's intent while taking into account contextual information retrieved from the Policy Information Point (PIP) in Step (iii). The contextual information may be information about facts or states of the environment or the system. For collecting contextual information in an automated manner, we assume that data subjects have smartphones equipped with some sensors for capturing environmental conditions. For instance, a smartphone can detect a fire alarm or an emergency situation such as a road accident. After collecting contextual information, an authorisation policy is populated and a data subject is asked about saving her preferences—in Step (iv)—for saving such an authorisation policy in order to authorise subsequent consent requests in an automated manner. Next, a consent response is sent to the data requester in Step (iv-b). Next, an authorisation policy is stored in the policy store in Step (v). Finally, this instantiation of authorisation policy is logged in Step (vi).

After an authorisation policy has been instantiated, any subsequent consent request will be received by the Policy Enforcement Point (PEP) running on the data subject's smartphone as we can see in Step (1). The PEP forwards this request to the Policy Decision Point (PDP) in Step (2). The PDP is responsible for fetching corresponding authorisation policies and collecting contextual information as we can see in Steps (3) and (4), respectively. Next, the PDP makes the decision by

evaluating authorisation policies against the request and contextual information provided. Then, it sends the consent response to the PEP in Step (5). The PEP also logs the decision made by the PDP in Step (6) and finally the consent response is sent to the data requester in Step (7).

We assume that the PEP and the administration point have access to the data subject's signing key that could be used for signing consent responses sent to the data requester. The consent log captures the complete details of actions that were taken by the data subject and decisions (e.g., signed consent responses) automatically made by the smartphones based on data subjects' intent. We note that a data subject has full access to her consent log.

The data subject has a right to update her consent policy, withdraw her consent by deactivating the consent policy or delete the policy altogether. In all these cases, data subjects have to interact with the administration point for any modification. Our architecture is flexible enough to cope with updates in the workflow of the healthcare providers or even in the law. All the healthcare providers have to do is to update TR and template policies stored in the TR and template store and delete, if any, existing authorisation policies managed by the policy store.

It is important to mention that data requesters can get access to the data in case of emergency using the *break glass* policy. In this case, the healthcare system could expect evidence of being in emergency situation. The healthcare system could defer verification of such evidence for the post-incident investigation.

14.5.1 Authorisation Policies

An authorisation policy specifies who is permitted (or denied) access to a resource under specific conditions. In ACTORS, an authorisation policy contains the following six fields:

1. **Data Requester Role:** It is role of the entity who makes the access request. It can contain either a single role or a set of roles.
2. **Data Requester ID:** It is ID of the one who makes the access request. Like the above field, this field can contain either a single ID or a list of IDs. This field is optional as permissions can be assigned to roles instead of specific IDs.
3. **Data Subject ID:** It refers to the data subject who owns the resources.
4. **Data Subject Resource:** It contains data subject resource(s) protected through the authorisation policy.
5. **Access Rights:** Access rights define the permission on the data subject resource.
6. **provided:** It contains a conditional expression that may contain a set of conditions combined with *and* and *or* logical operators. Each condition is a predicate that is bound to a variable. These variables can come from contextual information that may be facts or states about the system or the environment. The contextual information may include access purpose, access time, access date, data requester location and data subject location.

Fig. 14.4 An example of an authorisation policy

```

1 DataRequester.Role = {'Doctor' }
2 DataRequester.ID = {'Bob' }
3 DataSubject.ID = 'Alice'
4 DataSubject.Resource = {'Blood Test' }
5 AccessRights = {READ}
6 provided
7     (AccessPurpose = 'Diagnosis' or
8     AccessPurpose = 'Treatment') and
9     AccessTime ≥ 9:00

```

Figure 14.4 illustrates an example of an authorisation policy where Bob in a role doctor is permitted to have read access on Alice’s *Blood Test* report provided he makes the access request after 9:00 h for the purpose of diagnosis or treatment. The use of the Data Requester ID might seem redundant given the fact that the policy already has a Data Requester Role. However, it might be the case that the data subject might not want a specific requester to access her data. For instance, Alice does not want Eve (another doctor and Bob’s colleague) to read her *Blood Test* report. This requirement can be captured by specifying in the Data Requester ID the condition \neg ‘Eve’. The introduction of both positive and negative conditions could result in conflicting authorisation policies. For resolving conflicts, we can use existing resolution techniques, such as one proposed in [34].

We assume that once authorised as per authorisation policy, a data requester can access medical data of the patient for a certain time (say for the duration of the appointment). Once a time limit is reached, the data requester would not be able to access the data anymore.

14.5.2 Policy Templates

A policy template provides a structured format for instantiating authorisation policies on-the-fly. It is the authorisation policy specification with placeholders for variables that are assigned a value based on contextual information and a data subject’s intent. A data subject’s intent is about what a data subject can expect and can be captured based on actions taken by her. A policy template contains almost the same fields as an authorisation policy does. The fields of a very generic policy template are left blank so that they can be assigned a value based on contextual information. However, a list of options can be provided for each field. It means that a template field can only be filled, at the time of policy instantiation, with a value out of the list of options.

Figure 14.5 illustrates an example of a policy template. This policy template can be applied when a data requester is in role *Dentist* and the requested resource is *Dental Report* with access rights either *READ* or *WRITE* access and access purpose is either *Diagnosis* or *Treatment*. For rest of the fields, any value can be assigned based on contextual information and the data subject’s intent.

```

1 DataRequester.Role = {'Dentist'}
2 DataRequester.ID
3 DataSubject.ID
4 DataSubject.Resource = {'Dental Report'}
5 AccessRights = {READ, WRITE}
6 provided
7 AccessPurpose is 'Diagnosis' or 'Treatment'

```

Fig. 14.5 An example of a policy template

Generally, specifying an authorisation policy is difficult. However, policy templates, which could be provided by healthcare providers, make it easy for patients to instantiate required authorisation policies. For instantiation of authorisation policies from policy templates, a patient should be provided with usable interfaces with simple privacy controls. These privacy controls will lead to automatic generation of authorisation policies. Without loss of generality, our proposed architecture enables data subjects to update existing authorisation policies or create new ones.

Policy templates are associated with TR policies and goals that the TR policy is trying to achieve. For instance, the policy template in Fig. 14.5 can be applied when the goal of the patient is to visit a dentist. Therefore, such a template is associated with the TR policy managing that specific goal. Each TR policy can be associated with several templates. Based on contextual information and a data subject's intent, the TR policy can identify which policy template fulfils the criteria and then instantiates the required authorisation policy.

14.5.3 TR Policies

As already explained in Sect. 14.4, TR programs were introduced for continuously monitoring the behaviour of a robot while taking into account environmental changes. In ACTORS, we use TR policies for controlling the lifecycle of authorisation policies towards a specific goal, which is the management of data subject's consent in a given situation. Each TR policy might be associated with several policy templates from which authorisation policies can be instantiated. Several TR policies might be present on the data subject's smartphone. The selection of the appropriate TR policy is based on contextual information. The main advantage in using TR policies is that they provide a built-in prioritisation of actions needed for controlling the granting and revocation of data subjects' consent that reacts to the changes in the context in which the data subjects are interacting.

In the following section, we are going to provide details of how ACTORS can be used for the case study presented in Sect. 14.3.

14.6 Managing Consent in Healthcare Scenarios

ACTORS can be applied to any domain; however, we focus on healthcare scenarios as already described in Sect. 14.3, where consent needs to be captured and saved based on contextual information and the patient’s intent (note that in this context we assume that the patient is the data subject). For automatically instantiating authorisation policies regarding consent and managing lifecycle of those policies, we assume that each patient is provided a set of TR policies and policy templates at the time of registration with her healthcare provider. In fact, TR policies and policy templates are deployed on patients’ smartphone together with an application. Each TR policy can be associated with multiple policy templates. The smartphone application automatically selects the most appropriate TR policy and the policy template based on the consent request and contextual information. After instantiation of authorisation policies regarding consent, they are stored and enforced by the patient’s smartphone. It should be noted here that only policies and patient’s decisions are stored in the smartphone while the medical data is stored in the caregiver IT infrastructure. In this section, we explain in detail how we exploit the proposed approach, described in Sect. 14.5, for providing solutions for each scenario described in Sect. 14.3.

Patient Visiting Her GP In the scenario when a general practitioner (GP) needs the patient consent, a consent request is sent to the patient for providing access to a GP to requested resources. This consent request may be directly sent by the healthcare system to the patient when a GP makes an access request to the patient resources. This consent request may include information about the GP and the patient, the patient resources, an access purpose and access duration details. Based on the consent request together with contextual information, the most appropriate applicable TR policy and policy template are selected.

Figure 14.6 describes a TR policy that is applied when a GP needs a patient’s consent for accessing her data from his clinic. The name of this TR policy is *consentAtGPClinic* and *Patient* is the parameter. When the first consent request is made, consent is not available and the condition parts of rules at Line 3 and Line 5 evaluate to *false*. The condition part of rule at Line 7 also evaluates to *false* as no authorisation policy is instantiated yet, i.e., *instantiatedPolicy(Patient)* is *false*. However, the condition part of rule at Line 9 evaluates to *true*, so the action part of this rule is executed and the system waits for the patient decision for providing consent to her GP, i.e., *waitPatientDecision(Patient, GP)* is executed.

Once the patient provides consent for granting access to her GP on her resources, then *consentAvailable(Patient, GP)* becomes *true*. At the time of providing consent, a patient can be given an option to save her current preferences for providing her consent for similar consent requests when made in the same environment. If a patient does so, the condition part of rule at Line 3 becomes *true*; therefore, the authorisation policy regarding consent is instantiated from the policy template and then it is activated while at the same time, consent is sent.

```

1 tr policy consentAtGPClinic(Patient)
2
3 consentAvailable(Patient,GP) ∧ saveCurrentPreferences → instantiatePolicy(Patient) ⊗
   ↳activate(Patient.Policy) || sendConsent(Patient,GP)
4
5 consentAvailable(Patient,GP) → sendConsent(Patient,GP)
6
7 needsConsent(Patient,GP) ∧ instantiatedPolicy(Patient) ∧ ¬withdrawn(Patient.Policy) →
   ↳evaluatePolicy(Patient)
8
9 needsConsent(Patient,GP) → waitPatientDecision(Patient,GP)
10
11 deleteSavedPreferences(Patient) → remove(Patient.Policy)
12
13 activatePolicyRequest(Patient) → activate(Patient.Policy)
14
15 withdrawPolicyRequest(Patient) → withdraw(Patient.Policy)

```

Fig. 14.6 A TR policy for managing authorisation policy for providing consent to a GP

```

1 DataRequester.Role = {'GP'}
2 DataRequester.Name
3 DataSubject.Name
4 DataSubject.Resource
5 AccessRights
6 provided
7     AccessPurpose is 'Diagnosis' or 'Treatment'
8     AccessTime is within DutyHours
9     DataRequester.CurrentLocation = DataSubject.CurrentLocation
10    DataRequester.CurrentLocation = DataRequester.Clinic.Location

```

Fig. 14.7 A policy template for generating an authorisation policy for providing consent to a GP

Figure 14.7 illustrates a policy template that is applied when a patient visits her GP, as is evident from the data requester role that is GP only. The empty fields including data requester name, data subject name, data subject resource and access rights can be filled with values based on the consent request. However, there are certain conditions in the *provided* part of the policy template that are formulated at the time of instantiating an authorisation policy. These conditions include: the access purpose must be either *diagnosis* or *treatment*; access time must be in office hours; and both the patient and the GP must be present in the GP's clinic. These conditions are formulated based on contextual information that is collected from either patient's smartphone or the external information point, such as made available by the healthcare provider. The contextual information from a patient's smartphone may include information like patient's current location, while contextual information from the external information point may include information about location of GP's clinic and GP's duty hours. Once all the required information for the applicable policy template is retrieved, the authorisation policy is instantiated and activated.

```

1 DataRequester.Role = {'GP'}
2 DataRequester.ID = {'Bob'}
3 DataSubject.ID = 'Alice'
4 DataSubject.Resource = {'Blood Test'}
5 AccessRights = {READ}
6 provided
7   AccessPurpose = 'Diagnosis' and
8     (AccessTime ≥ 9:00 and AccessTime ≤ 17:00) and
9     DataSubject.CurrentLocation = 'Milan' and
10    DataRequester.CurrentLocation = 'Milan'

```

Fig. 14.8 An authorisation policy for providing consent to a GP

Figure 14.8 shows the instantiated authorisation policy regarding consent, expressing that a GP Bob can get patient Alice's consent for *READ* access on Alice's *Blood Test* when accessed for the *Diagnosis* purpose during the duty hours (that is, between 9:00 and 17:00 h) from Bob's clinic located in *Milan*.

A patient may decide to withdraw her consent. In this case, the condition part of rule at Line 15, i.e., condition *withdrawPolicyRequest(Patient)*, becomes *true* and the authorisation policy is withdrawn by invoking *withdraw(Patient.Policy)* function. Furthermore, a patient can decide to activate her withdrawn consent. In this case, condition *activatePolicyRequest(Patient)* becomes *true* and *activate(Patient.Policy)* function is invoked for activating the authorisation policy. Last but not least, a patient may also choose to delete forever her saved preferences for automatically providing consent. In this case, *deleteSavedPreferences(Patient)* becomes *true* and *remove(Patient.Policy)* function is invoked for deleting the instantiated authorisation policy.

In case if a GP needs the patient consent when the patient has already saved preferences for providing consent automatically to her GP and consent is not withdrawn yet then consent will be provided after evaluating the consent request and contextual information against the instantiated authorisation policy, see rule in Fig. 14.6 at Line 7. We assume that the consent request is same as already described above. However, we have to collect contextual information in order to evaluate the authorisation policy for providing consent. The patient's smartphone may provide information about her location and the current time while the information about the GP's location can be collected from the external information point. This may be the healthcare system or the GP's smartphone which may provide GP's location information to the patient's smartphone. Based on the consent request and contextual information, the authorisation policy is evaluated (see rule in Fig. 14.6 at Line 7). After the evaluation of the authorisation policy, consent becomes available and the consent response is automatically sent by the patient's smartphone (see rule in Fig. 14.6 at Line 5). The consent response contains patient consent if the authorisation policy evaluates to *true*, otherwise it may contain an error message.

A patient may decide not to save her current preferences for providing consent automatically to her GP. In such a case, the patient will be explicitly asked each time (see rule in Fig. 14.6 at Line 9) and consent will be provided once the patient takes her decision (see rule in Fig. 14.6 at Line 5).

Patient Visiting a Cardiologist A cardiologist may also need the patient's consent while accessing the patient resources. Like the above scenario, a patient receives the consent request. This consent request may include information about the cardiologist and the patient, the patient resources, an access purpose and access duration details. The additional point in this scenario, as compared to the previous scenario, is that a cardiologist is provided consent for getting access on the patient resources as long as the treatment may last. In other words, the saved preferences for providing consent are deleted automatically right after the treatment.

Figure 14.9 shows the TR policy (*consentAtSpecialistClinic*) for managing authorisation to provide consent to a specialist. The TR policy is similar to one in Fig. 14.6. In the case of a cardiologist, the TR policy of specialist is selected. Observe that this TR policy is very generic and can be applied to other specialists, such as a dentist and a gynaecologist. However, there is a specific policy template for each specialist. The policy template for cardiologist is shown in Fig. 14.10 and is restricted to only resources that could be accessed by a cardiologist, such as: *ECG Report*, *Cardiography*, and *Engyography*. This is different from the policy template of the above scenario, as the resource field in Fig. 14.7 is left empty, indicating that a GP can obtain consent to access any resource.

Figure 14.11 shows the authorisation policy regarding consent for a cardiologist when a patient intends to save her preferences until she is treated. The authorisation policy expresses that a cardiologist David can get patient Alice's consent for *READ* and *WRITE* access on Alice's *ECG Report* when accessed for *Diagnosis* purpose during the duty hours (that is, between 9:00 and 17:00 h) from David's clinic that is located in *Como*.

```

1 tr policy consentAtSpecialistClinic(Patient)
2
3 consentAvailable(Patient,Specialist) ∧ saveCurrentPreferences → instantiatePolicy(Patient)
   ↳ ⊗ activate(Patient.Policy) || sendConsent(Patient,Specialist)
4
5 consentAvailable(Patient,Specialist) → sendConsent(Patient,Specialist)
6
7 needsConsent(Patient,Specialist) ∧ instantiatedPolicy(Patient) ∧
   ↳¬withdrawn(Patient.Policy) → evaluatePolicy(Patient)
8
9 needsConsent(Patient,Specialist) → waitPatientDecision(Patient,Specialist)
10
11 timeout(Patient.Policy) ∨ deleteSavedPreferences(Patient) → remove(Patient.Policy)
12
13 activatePolicyRequest(Patient) → activate(Patient.Policy)
14
15 withdrawPolicyRequest(Patient) → withdraw(Patient.Policy)

```

Fig. 14.9 A TR policy for providing consent to a specialist

```

1 DataRequester.Role = {'Cardiologist'}
2 DataRequester.ID
3 DataSubject.ID
4 DataSubject.Resource = {'ECG Report', 'Cardiography', '
    ↪Engyography'}
5 AccessRights = {READ, WRITE}
6 provided
7     AccessPurpose is 'Diagnosis' or 'Treatment'
8     AccessTime is within DutyHours
9     DataRequester.CurrentLocation = DataSubject.CurrentLocation
10    DataRequester.CurrentLocation = DataRequester.Clinic.Location

```

Fig. 14.10 A policy template for generating an authorisation policy for providing consent to a cardiologist

```

1 DataRequester.Role = {'Cardiologist'}
2 DataRequester.ID = {'David'}
3 DataSubject.ID = 'Alice'
4 DataSubject.Resource = {'ECG Report'}
5 AccessRights = {READ, WRITE}
6 provided
7     AccessPurpose = 'Diagnosis' and
8         (AccessTime  $\geq$  9:00 and AccessTime  $\leq$  17:00) and
9         DataSubject.CurrentLocation = 'Como' and
10        DataRequester.CurrentLocation = 'Como'

```

Fig. 14.11 An authorisation policy for providing consent to a cardiologist

The authorisation policy regarding consent for a cardiologist may automatically be deleted once the treatment completes. This information about treatment duration can be collected by the patient at the time of saving her preferences. For instance, it may be included in the consent request or can be collected as contextual information from the information point made available by the service provider. Once the treatment duration expires (starting from when the first consent request is made), condition *timeout(Patient.Policy)* becomes automatically *true* and *remove(Patient.Policy)* function is invoked for deleting the instantiated authorisation policy according to the rule at Line 11 in Fig. 14.9. Alternatively, a patient may decide to delete her saved preferences during the treatment duration as already considered in above scenario.

Patient in an Emergency Situation In an emergency situation, the emergency response team may need a patient's consent in order to get an access to her medical data for the treatment purpose. Similar to above scenarios, the patient receives the consent request, which may include information about the emergency response team, the patient resources, an access purpose and access duration details. Similar to the cardiologist scenario, we consider that the patient intends to provide her consent as long as the treatment may last. Technically, the saved preferences for providing consent are deleted automatically right after the treatment. The TR policy for specialist, shown in Fig. 14.9, can also be applied for this scenario.

```

1 DataRequester.Role = {'EmergencyResponseTeam'}
2 DataRequester.Name
3 DataSubject.Name
4 DataSubject.Resource = {'Allergy Report', 'Blood Test'}
5 AccessRights = {READ}
6 provided
7     There is an Emergency situation
8     AccessPurpose is 'Diagnosis' or 'Treatment'
9     DataRequester.CurrentLocation = DataSubject.CurrentLocation

```

Fig. 14.12 A policy template for generating an authorisation policy for providing consent to the emergency response team

```

1 DataRequester.Role = {'EmergencyResponseTeam'}
2 DataRequester.ID = {'Fayne'}
3 DataSubject.ID = 'Alice'
4 DataSubject.Resource = {'Allergy Report'}
5 AccessRights = {READ}
6 provided
7     Emergency = TRUE and
8     AccessPurpose = 'Diagnosis' and
9     DataSubject.CurrentLocation = 'Aachen' and
10    DataRequester.CurrentLocation = 'Aachen'

```

Fig. 14.13 An authorisation policy for providing consent to the emergency response team

The policy template applied in emergency situation is shown in Fig. 14.12. In the *provided* part of the policy template for emergency situations, we include the condition for capturing the notion of emergency situation, i.e., *There is an Emergency situation*. Furthermore, we omit also the condition *AccessTime is within DutyHours*, in contrast to the policy template for a GP shown in Fig. 14.7, considering the fact that the emergency can happen at any time. For restraining access in emergency situations, the resource field of the policy template is set to *Allergy Report* and *Blood Test*. Moreover, we consider *READ* only access in emergency situations.

Figure 14.13 shows the authorisation policy for providing consent to the emergency response team. This authorisation policy is instantiated when an emergency occurs in *Aachen* and *Fayne*, a member of the emergency response team, requests *READ* access on (a patient) Alice's *Allergy Report* for *diagnosis*, while Alice provides her consent and also saves her preferences for subsequent requests issued in the same environment. The occurrence of an emergency situation may be detected using a patient's smartphone.

There are few important points to be considered. First, we are instantiating one authorisation policy per instance of the emergency response team. Alternatively, it may also be possible to instantiate the authorisation policy at the role level (i.e., *EmergencyResponseTeam*) instead of at the instance level (i.e., *Fayne*). Second, the patient may be in the unconscious state and may not be able to provider her consent. In such situations, authorisation policies can be instantiated from break-the-glass

policy templates, without asking patients. In other words, the emergency response team may provide consent on patient's behalf when patients are in the unconscious state. Here, the unconsciousness state can be incorporated at the time of sending the consent request by members of the emergency response team to the patient's smartphone. Finally, as an ultimate break-the-glass policy in case the smartphone is not reachable or not functioning, the emergency team can specify the current circumstances together with the request for accessing the patient's medical data. This information can subsequently be checked in a post-incident analysis to make sure that this access mode is not abused. Again, it should be underlined here that the medical data are not stored in the smartphone device.

14.7 Related Work

In [1], Aboelfotoh et al. propose a mobile-based architecture for integrating Personal Health Record (PHR), where allows patients to control their data through their mobile devices. To manage the lifecycle of consent, they use the goal-driven approach, proposed by Asghar and Russello in [3]. In addition, they comply with the privacy consent direction [15] of Health Level 7 (HL7) [10], a reference guide for exchanging healthcare data.

Curren and Kaye [8] provide a legal background that signifies importance of consent withdrawal and revocation. According to their analysis, implementing consent withdrawal and revocation is not straightforward in practice, in particular when we address all the related legal complications. In [32], Pruski introduces e-CRL, a language for expressing patients' consent to regulate access to their health information.

Russello et al. [35] propose a consent-based framework that enables patients to control disclosure of their medical data, where the mechanism of capturing consent is integrated with workflows. The idea is to automatically generate Ponder2 style of authorisation policies [9] that depend on workflows. However, there is no automatic mechanism for managing the lifecycle of consent, such as consent withdrawal, activation or deletion. Asghar and Russello [2] suggest a mechanism for managing the consent lifecycle. They introduce a notion of very expressive consent represented as a consent policy. However, they assume that a data subject defines his/her consent policies; unfortunately, such a solution may not be acceptable because data subjects may not be able to understand low-level policy details.

Wuyts et al. [43] incorporate patient consent to healthcare systems. They use the XACML policy language for defining access control on medical data and retrieve consent from the Policy Information Point (PIP). They express consent as a set of pre-defined attributes and store it in the database. A similar approach is used in [18], which is an authorisation framework for sharing EHR. The main issue with both approaches is that the set of pre-defined attributes may not be sufficient to capture consent, as it may involve certain conditions. To overcome this issue, there are approaches [2, 31] in which consent is treated as an authorisation policy; however,

this raises some other problems. First, this approach requires users to specify low-level details, which a normal user may not be aware of, at the time of policy creation. Second, there is no automatic mechanism for managing the consent lifecycle.

EnCoRe [27, 42] aims at managing consent of users in order to regulate access to their personal data. In EnCoRe, a user is expected to define her preferences regarding consent, which are stored by enterprises. Once any piece of personal data is requested, these preferences are checked by the enterprises before granting access to the requested data. However, it may be cumbersome for users to define such complicated preferences. In our proposed solution, users' consent can be captured and managed dynamically by taking into account contextual information. Furthermore, our proposed approach offers more control and access to users as consent is stored and managed on their smartphone.

Luger and Rodden [21, 22] advocate how consent is a critical concern in pervasive computing. They describe issues with existing consent systems and highlight challenges and recommendations for a consent management system.

Marinovic et al. [24] employ TR policies for continuously monitoring the nursing home, where caregivers (including nurses, head-nurses, patients and students) are equipped with mobile devices for running their corresponding TR policies. They use TR policies to manage all activities of a caregiver using one workflow specification while we use TR policies with the goal of capturing consent that may involve instantiation of authorisation policies regarding consent and management of their lifecycle, consisting withdrawal and activation of consent.

Illner et al. [16, 17] suggest an automated approach for managing services related to distributed and embedded systems in dynamic environments. In their approach, various configurations for the services are generated and mapped to specific environmental conditions only once at the design time when system is setup while appropriate configurations for the services are activated at runtime when certain environmental conditions hold. The shortcoming of this approach is that the configurations are defined statically while our goal-based approach is dynamic in a sense that authorisation policies do not need to be specified in advance and are instantiated automatically while taking into account environmental conditions.

Johnson et al. [19] suggest a general approach for creating policy templates. A policy template provides users with a structured format for authoring policies. In our proposed solution, a healthcare provider may consider this work for generating policy templates. Chan and Kwok [4] describe a method to create policies automatically based on observed events. They use the Singular Value Decomposition (SVD) technique for modelling correlation between events and policies and then create new policies or select recommended policies based on the correlation. Unfortunately, the SVD technique may not always choose the fine-grained policies while our proposed approach always generates the fine-grained authorisation policies based on environmental conditions.

Fu et al. [13, 14] propose how to automatically generate required IPSec policies without manual configuration. The idea is to define high-level security requirements and then automatically generate a set of IPSec policies that can satisfy all security requirements. The main problem is that this approach incurs high performance

overhead for finding the required set of policies as the proposed algorithm needs to go through a large number of possibilities before halting. Instead of generating a set of authorisation policies, our proposed approach generates only a single authorisation policy while taking into account contextual information and user intent.

14.8 Conclusion and Future Work

With the increasing attention towards the notion of data subjects' consent to be integrated in access control mechanisms, the task of appropriately capturing security requirements in policy specification is becoming daunting. This increases the risk of introducing errors in the policy specification that might compromise the privacy of medical data. In light of this, in this chapter we proposed ACTORS, a goal-driven approach where authorisation policies are managed by TR policies that aim to capture the consent preferences of a data subject. As we have shown, in our scenario data subjects might want to handle consent in accordance with the actual situation and context. TR policies are structured in such a way that rules at the top are closer to the goal of the policy, while rules at the bottom are more relevant when the goal is not close to be achieved. This is very natural for humans to grasp; therefore, a security administrator can capture more naturally the security requirements.

As future work, we are planning to focus on securing the mobile device where the consent application is installed. As mobile device could be stolen or misplaced, there is a need to make sure that the data subject does not loose control over her data and consent. We are exploring several approaches including dynamic authentication mechanisms to authenticate users in a seamless manner, i.e., without requiring unnecessary interactions with the device.

Another area that deserves investigation regards the enforcement of cross-domain policies. In this setting, it is difficult for the security administrator to have all the details of the different domains in which the data of the user might end up. Our idea is to establish a mapping of the policy templates from one domain to the other, say by means of ontologies.

We are planning to perform a thorough evaluation with the medical school in our university. Our experience so far in capturing security requirements with the use TR policies, is very positive. ACTOR has already captured the requirements of one of the testbeds in the ENDORSE project. Another interesting area would be capturing consent for handling personal data of customers of a commercial entity. We can apply the same concept to solve other real-world problems. Terms and conditions at signing up or installing a software and granting app permissions in Android (or approving app restrictions in iOS) are few interesting problems among many others.

References

1. Aboelfotoh, M., Martin, P., Hassanein, H.: A mobile-based architecture for integrating personal health record data. In: IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom), 2014, pp. 269–274 (2014)
2. Asghar, M., Russello, G.: Flexible and dynamic consent-capturing. In: Camenisch, J., Kessoglu, D. (eds.) Open Problems in Network Security. Lecture Notes in Computer Science, vol. 7039, pp. 119–131. Springer, Berlin (2012)
3. Asghar, M.R., Russello, G.: ACTORS: A goal-driven approach for capturing and managing consent in e-health systems. In: 2012 IEEE International Symposium on Policies for Distributed Systems and Networks (POLICY), pp. 61–69 (2012)
4. Chan, H., Kwok, T.: A policy-based management system with automatic policy selection and creation capabilities by using a singular value decomposition technique. In: Seventh IEEE International Workshop on Policies for Distributed Systems and Networks, 2006. Policy 2006, pp. 96–99 (2006)
5. Clarke, R.: econsent: A critical element of trust in ebusiness. In: BLED 2002 Proceedings, p. 12 (2002)
6. Coiera, E., Clarke, R.: e-consent: the design and implementation of consumer consent mechanisms in an electronic environment. *J. Am. Med. Inform. Assoc.* **11**(2), 129–140 (2004)
7. Communities, E.: Directive 1999/93/EC of the european parliament and of the council of 13 december 1999 on a community framework for electronic signatures (1999). <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1999L0093:20081211:EN:PDF>
8. Curren, L., Kaye, J.: Revoking consent: a “blind spot” in data protection law? *Comput. Law Secur. Rev.* **26**(3), 273–283 (2010)
9. Damianou, N., Dulay, N., Lupu, E., Sloman, M.: The ponder policy specification language. In: Sloman, M., Lupu, E., Lobo, J. (eds.) Policies for Distributed Systems and Networks. Lecture Notes in Computer Science, vol. 1995, pp. 18–38. Springer, Berlin (2001)
10. Dolin, R.H., Alschuler, L., Boyer, S., Beebe, C., Behlen, F.M., Biron, P.V., Shvo, A.S.: HI7 clinical document architecture, release 2. *J. Am. Med. Inform. Assoc.* **13**(1), 30–39 (2006)
11. Earp, J.B., He, Q., Stufflebeam, W., Bolchini, D., Jensen, C., et al.: Financial privacy policies and the need for standardization. *IEEE Secur. Priv.* **2**(2), 36–45 (2004)
12. European Communities: Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, year=1995, howpublished = http://ec.europa.eu/justice/policies/privacy/docs/95-46-ce/dir1995-46_part1_en.pdf
13. Fu, Z.: Network management and intrusion detection for quality of network services. Ph.D.in Computer Science, North Carolina State University (2001)
14. Fu, Z.J., Wu, S.F.: Automatic generation of IPSec/VPN security policies in an intra-domain environment. In: 12th International Workshop on Distributed Systems: Operations & Management (2001)
15. Health Level Seven International: HI7 implementation guide for cda release 2: Privacy consent directives, release 1. http://gforge.hl7.org/gf/download/frsrelease/977/10295/CDAR2_IG_CONSENTDIR_R1_N1_2013MAY.pdf (2013)
16. Illner, S., Krumm, H., Pohl, A., Luck, I., Manka, D., Sparenberg, T.: Policy controlled automated management of distributed and embedded service systems. In: Parallel and Distributed Computing and Networks, pp. 710–715 (2005)
17. Illner, S., Pohl, A., Krumm, H., Luck, I., Manka, D., Sparenberg, T.: Automated runtime management of embedded service systems based on design-time modeling and model transformation. In: 2005 3rd IEEE International Conference on Industrial Informatics, INDIN '05, pp. 134–139 (2005)
18. Jin, J., Ahn, G.J., Hu, H., Covington, M.J., Zhang, X.: Patient-centric authorization framework for sharing electronic health records. In: Proceedings of the 14th ACM Symposium on Access Control Models and Technologies, SACMAT '09, pp. 125–134. ACM, New York, NY (2009)

19. Johnson, M., Karat, J., Karat, C., Grueneberg, K.: Usable policy template authoring for iterative policy refinement. In: 2010 IEEE International Symposium on Policies for Distributed Systems and Networks (POLICY), pp. 18–21 (2010)
20. Lawson, P., O'Donoghue, M.: Approaches to consent in Canadian data protection law. In: Lessons from the Identity Trail: Anonymity, Privacy and Identity in a Networked Society, pp. 23–42 (2009) <https://goo.gl/VqPUwF>
21. Luger, E., Rodden, T.: An informed view on consent for ubicomp. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '13, pp. 529–538. ACM, New York, NY (2013)
22. Luger, E., Rodden, T.: Terms of agreement: rethinking consent for pervasive computing. *Interact. Comput.* **25**(3), 229–241 (2013) doi:10.1093/iwc/iws017
23. Malone, P., McLaughlin, M., Leenes, R., Ferronato, P., Lockett, N., Guillen, P.B., Heistracher, T., Russello, G.: ENDORSE: a legal technical framework for privacy preserving data management. In: Proceedings of the 2010 Workshop on Governance of Technology, Information and Policies, pp. 27–34. ACM (2010)
24. Marinovic, S., Twidle, K., Dulay, N., Sloman, M.: Teleo-reactive policies for managing human-centric pervasive services. In: Network and Service Management (CNSM), 2010 International Conference on, pp. 80–87 (2010)
25. McDonald, A.M., Cranor, L.F.: Cost of reading privacy policies, the. *ISJLP* **4**, 543 (2008)
26. McNair, L., Costello, A.: Electronic informed consent: a new industry standard (2014) http://www.wegclinical.com/wp-content/uploads/2014/03/eConsent-White-Paper_FINAL.pdf
27. Mont, M.C., Pearson, S., Kounga, G., Shen, Y., Bramhall, P.: On the management of consent and revocation in enterprises: setting the context. HP Laboratories, Technical Report HPL-2009-49 (2009)
28. Nilsson, N.J.: Teleo-reactive programs for agent control. *J. Artif. Intell. Res.* **1**, 139–158 (1994)
29. Nissenbaum, H.: Privacy in context: technology, policy, and the integrity of social life. Stanford University Press, Stanford (2009)
30. OASIS Standard: eXtensible Access Control Markup Language (XACML) Version 3.0. <http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.pdf> (2013)
31. O'Keefe, C.M., Greenfield, P., Goodchild, A.: A decentralised approach to electronic consent and health information access control. *J. Res. Pract. Inf. Technol.* **37**(2), 161–178 (2005)
32. Pruski, C.: e-CRL: A rule-based language for expressing patient electronic consent. In: Second International Conference on eHealth, Telemedicine, and Social Medicine, 2010. ETELEMED '10, pp. 141–146 (2010)
33. Report of the Secretary's advisory committee on automated personal data systems. U.S. Department of Health, Education & Welfare, Records, Computers, and the Rights of Citizens (1973)
34. Russello, G., Dong, C., Dulay, N.: Authorisation and conflict resolution for hierarchical domains. In: Eighth IEEE International Workshop on Policies for Distributed Systems and Networks, 2007. POLICY '07, pp. 201–210 (2007)
35. Russello, G., Dong, C., Dulay, N.: Consent-based workflows for healthcare management. In: IEEE Workshop on Policies for Distributed Systems and Networks, 2008. POLICY 2008, pp. 153–161 (2008)
36. Saltzer, J., Schroeder, M.: The protection of information in computer systems. *Proc. IEEE* **63**(9), 1278–1308 (1975)
37. Schwartz, P.M.: The eu-us privacy collision: a turn to institutions and procedures (2013)
38. Schwartz, P.M., Solove, D.J.: The PII problem: privacy and a new concept of personally identifiable information. *NYUL Rev.* **86**, 1814 (2011)
39. Solove, D.J.: Introduction: Privacy self-management and the consent dilemma. *Harv. Law Rev.* **126**, 1880 (2012)
40. Turow, J., Feldman, L., Meltzer, K.: Open to exploitation: America's shoppers online and offline (2005)

41. Twidle, K., Dulay, N., Lupu, E., Sloman, M.: Ponder2: a policy system for autonomous pervasive environments. In: International Conference on Autonomic and Autonomous Systems, pp. 330–335 (2009)
42. Whitley, E.A.: Informational privacy, consent and the “control” of personal data. *Inf. Secur. Tech. Rep.* **14**(3), 154–159 (2009)
43. Wuyts, K., Scandariato, R., Verhenneman, G., Joosen, W.: Integrating patient consent in e-Health access control. *Int. J. Secure Softw. Eng. IGI Global* **2**(2), 1–24 (2011). Partner: KUL; project: NESSoS
44. Zhou, X., Demetriadou, S., He, D., Naveed, M., Pan, X., Wang, X., Gunter, C.A., Nahrstedt, K.: Identity, location, disease and more: inferring your secrets from android public resources. In: Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, pp. 1017–1028. ACM, New York (2013)

Chapter 15

e-Health Cloud: Privacy Concerns and Mitigation Strategies

Assad Abbas and Samee U. Khan

15.1 Introduction

Technological developments have greatly influenced conventional healthcare practices. The healthcare sector has advanced from conventional clinical settings with paper-based medical prescriptions to Electronic Medical Records (EMR), Personal Health Records (PHR), and Electronic Health Records (EHR) [60]. The need to integrate electronic medical information from dispersed locations, such as clinics, hospitals, clinical laboratories, and health insurance organizations, has given rise to the phenomenon called *e-Health*. The World Health Organization (WHO) defines e-Health as “*the transfer of healthcare information and resources to healthcare professionals and consumers by employing the Information Technology (IT) infrastructure and e-commerce practices*” [53]. However, the exchange and integration of electronic medical information, managed by several healthcare providers and other participating organizations, is inflated and difficult to manage, which calls for utilizing the cloud computing services in the healthcare domain [60]. The cloud computing model has relieved healthcare organizations of the strenuous tasks of infrastructure management and has urged them to become accustomed to third-party IT service providers [2]. Moreover, the cloud computing paradigm has exhibited great potential: (a) *to enhance collaboration among various participating entities to the healthcare domain* [6], and (b) *to offer the most anticipated benefits, such as scalability, agility, cost effectiveness, and round the clock availability of health-related information* [3, 54].

On the other hand, due to the sharing and storage of sensitive electronic health data and Personal Health Information (PHI) through Internet, various privacy

A. Abbas (✉) • S.U. Khan
Department of Electrical and Computer Engineering,
North Dakota State University, Fargo, ND, USA
e-mail: assad.abbas@ndsu.edu; samee.khan@ndsu.edu

and security concerns arise [28]. The literature pertaining to the e-Health clouds discusses the apprehensions about the probable disclosure of health information to entities that are not supposed to have access. One of the key reasons for patients' concerns about the privacy of PHI is the distributed architecture of the cloud. The storage of gigantic volumes of confidential health data to third-party data centers and the transmission of these data over networks is vulnerable to disclosure or theft [22]. Particularly, in public clouds, administered by commercial service providers, health data privacy is the most anticipated concern. Therefore, the Cloud Service Providers (CSP) should not only identify but also deal with health data security issues to maximize the trust level of patients and healthcare organizations [1]. Governments have also shown interest to protect the privacy of health data. For example, in the United States, the use and disclosure of patient health information is protected by the Health Insurance Portability and Accountability Act (HIPAA). The health data privacy rules specified by HIPAA offer federal protection for the PHI and ensure the confidentiality, integrity, and availability of electronic health information [48]. Likewise, the Health Information Technology for Economic and Clinical Health (HITECH) Act [20] also mandates the secure exchange of electronic health information.

Various approaches, such as cryptographic and non-cryptographic methods, are used to preserve the privacy of health data in the cloud. The majority of the solutions use certain cryptographic techniques to conceal the contents of health records, while quite a few solutions, such as [15, 40, 54] are based on non-cryptographic approaches, using policy-based authorizations. The benefit of cryptographic techniques is that they not only are capable of encrypting the data in storage and over the network [24], but they also employ authentication mechanisms requiring decryption keys and verification through digital signatures. Moreover, fine-grained and patient-centric access control mechanisms have been deployed that enable patients to specify the individuals who could have access to health data. Furthermore, quite a few privacy preserving solutions allow the patients themselves to encrypt the health data and provide the decryption keys to the individuals with right-to-know privilege.

This chapter encompasses the recent efforts that have been made to preserve the privacy of the health data in the cloud environment. We highlight the threats to the health data in the cloud and present a discussion on the requirements to be fulfilled in order to mitigate these threats. Moreover, we discuss the benefits of cloud computing and the cloud deployment models in the context of healthcare. Furthermore, the strengths and weaknesses of each of the discussed strategy to preserve privacy are reported and some open research issues are also highlighted.

The chapter is organized as follows. Section 15.2 presents an overview of the preliminary concepts of cloud computing in terms of healthcare. Section 15.3 presents the recent strategies developed to overcome the privacy issues of health. Section 15.4 presents discussion on the performance of discussed strategies and highlights open issues, and Sect. 15.5 concludes the chapter.

15.2 An Overview of the e-Health Cloud

The e-Health cloud can be regarded as a standard platform that offers standardized services to manage large volumes of health data [19]. The e-Health cloud ensures the service provision for storage and processing of different types of health records that are originated and utilized by multiple providers and other participating entities, such as pharmacies, laboratories, and insurance providers. Typically, the health records in an e-Health system include the EMRs, the EHRs, and the PHRs. Each of the aforementioned type of health records is the electronic version of patient health information. However, there are certain differences that should be understood.

The EMR is the electronic version of a patient's health information that is created, used, and maintained by the healthcare providers or care delivery organizations. The EMRs contain information about the diagnosis obtained through the clinical decision support system, clinical notes, and medication. The EHRs, on the other hand, present a broader view of the patients' health information. A subset of the information contained in the EMRs is also present in the EHRs. However, the EHRs are shared for the purpose of consultation and treatment among multiple healthcare providers belonging to different care delivery organizations [60]. The PHRs are the health records that are managed by the patients themselves and comprise of the information instigated from diverse sources. The typical information that a PHR may contain includes treatments and diagnosis, surgeries, laboratory reports, insurance claims data, as well as personal notes and wellness charts that the patients use in order to keep track of their health [26]. Figure 15.1 presents a distinction among the EMRs, the EHRs, and the PHRs.

15.2.1 *e-Health Cloud Benefits and Opportunities*

Cloud computing, besides various other scientific and business domains, has greatly impacted the healthcare sector. Shifting the health records to the cloud environment brings many opportunities and benefits to the health service providers, as discussed in the following sections.

15.2.1.1 Cost Reduction

Cloud computing relieves the organization of the obligations to purchase the hardware and software to manage their data [41]. Instead, the clients pay to the CSPs for the services on the basis of pay-as-you-go model. Adopting cloud services eradicates the need for possessing the aforesaid resources. Therefore, one of the key incentives for healthcare organizations to embrace cloud services is the significant cost reduction in terms of purchase of computing resources.

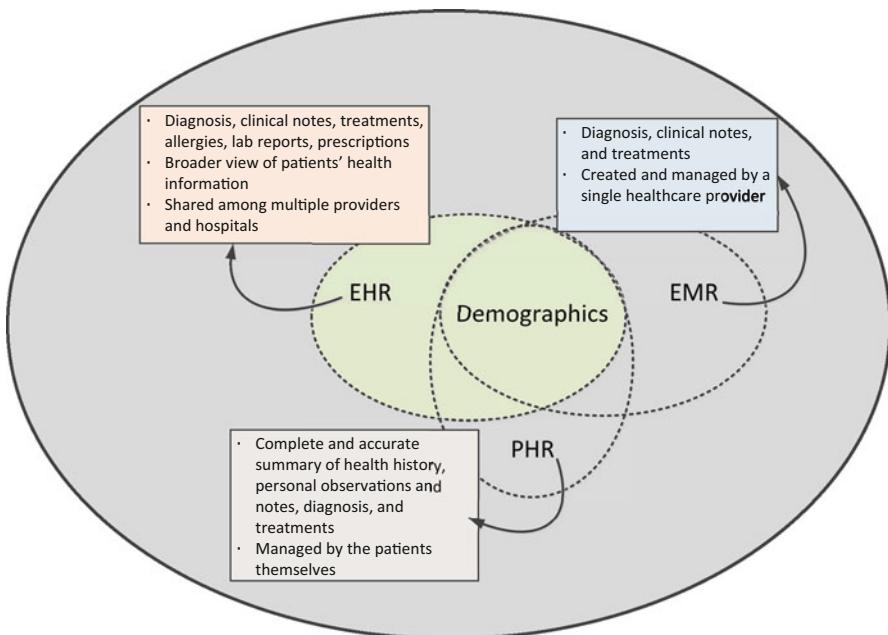


Fig. 15.1 Distinction among the EMR, PHR, and EHR

15.2.1.2 Easy Infrastructure Management

For healthcare organizations with limited hardware and software resources, and support staff with inadequate technical capabilities, the tasks of infrastructure management can be cumbersome. Therefore, healthcare organizations can avoid the arduous management tasks by delegating them to cloud service providers [25].

15.2.1.3 Availability

Contemporary developments in healthcare systems, focusing on accessing information anytime and anywhere, offer health service providers with a great opportunity to move healthcare information to the cloud. This will also ensure the ubiquity of the health information for all of the stakeholders of the e-Health cloud.

15.2.1.4 Scalable Healthcare Services

Scalability refers to expanding the IT infrastructure by increasing the number of computers, network interconnections, and storage capacity of the data centers while maintaining performance [41]. The latest trends in healthcare demand the

scalability of health cloud infrastructure to facilitate all the geographically scattered stakeholders, such as patients, doctors, clinical staff, lab staff, and insurance companies. Therefore, cloud computing has the ability to facilitate large numbers of healthcare providers with millions of health records.

15.2.2 Deployment Models for Cloud Based e-Health Systems

To offer the cloud computing services for healthcare, mostly the following deployment models are used: (a) private cloud, (b) public cloud, and (c) hybrid cloud. The cloud deployment models for the healthcare domain work in the same way as in other business and scientific domains.

15.2.2.1 Private Cloud

The private cloud is usually owned and administered by an organization [42, 67]. In a particular e-Health scenario, the cloud infrastructure, such as the storage and processing units, are typically managed by the hospitals or any designated third-party. However, due to the restricted exposure to the public Internet, the EMRs stored at the private cloud are considered much secure as compared to the other deployment models. This is because the EMRs in a private cloud environment are only accessed by the employees of the healthcare organizations, who are generally considered as trusted (with few exceptions). A private e-Health cloud is illustrated in Fig. 15.2.

15.2.2.2 Public Cloud

The public cloud consists of the shared physical infrastructure that is managed by the third-party providers [42]. The organizations utilizing the cloud services procure the services from the Cloud Service Providers (CSPs). In a public e-Health cloud, the EHRs may be shared among different participating entities, such as clinics, hospitals, insurance companies, pharmacies, and clinical laboratories. Moreover, the EHRs are stored at the off-premises servers managed by the CSPs. Therefore, the EHRs are highly vulnerable to malicious attacks and forgery attempts both by the internal, as well as external, entities. Consequently, mechanisms are needed to mitigate the privacy concerns and to ensure confidentiality through strong cryptographic techniques, patient-centric access control, and efficient signature verification schemes. An illustration of the public cloud in the context of e-Health is presented in Fig. 15.3.

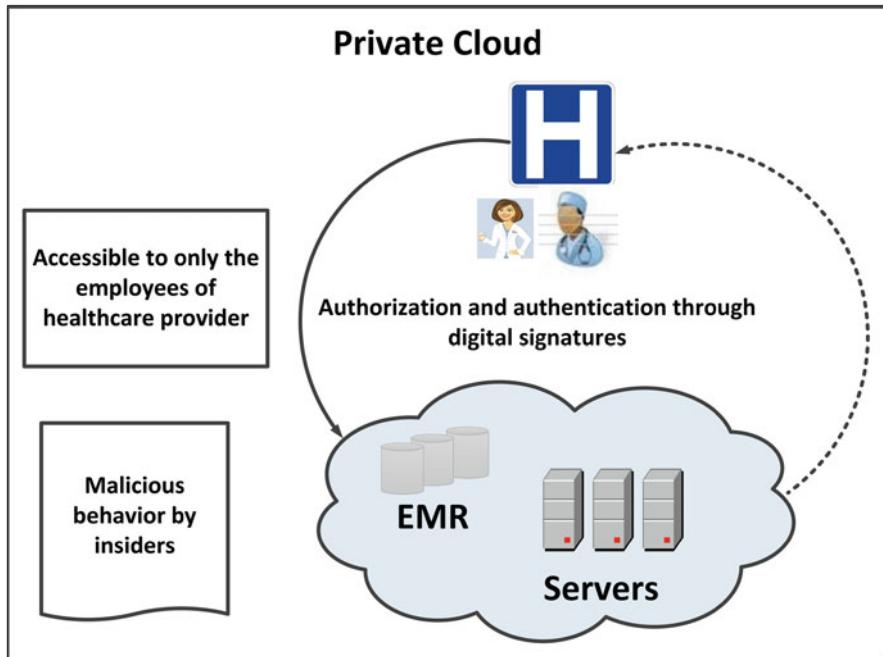


Fig. 15.2 An illustration of a private cloud in context of e-Health

15.2.2.3 Hybrid Cloud

The hybrid clouds are a combination of two or more cloud providers (public or private) such that each of them operates independently but are bound together through standardized technologies [67]. The hybrid cloud deployment model is truly beneficial for healthcare services, where the healthcare providers with limited physical resources and interested in using their legacy systems can procure the third-party services to store huge clinical and medical imaging data [49]. However, a key limitation of this model is that it requires more security measures in order to achieve privacy. An illustration of a hybrid cloud is presented in Fig. 15.4.

15.2.3 Threats to Health Data Privacy in the Cloud

The PHI in transit or at a data center may contain sensitive health data, such as patient medical histories, current disease information, symptoms, and various laboratory test reports. The security and privacy of health data may be at stake in a variety of ways. As an example, health data may be susceptible to access by unauthorized external entities when stored at a CSP or in transit from a general

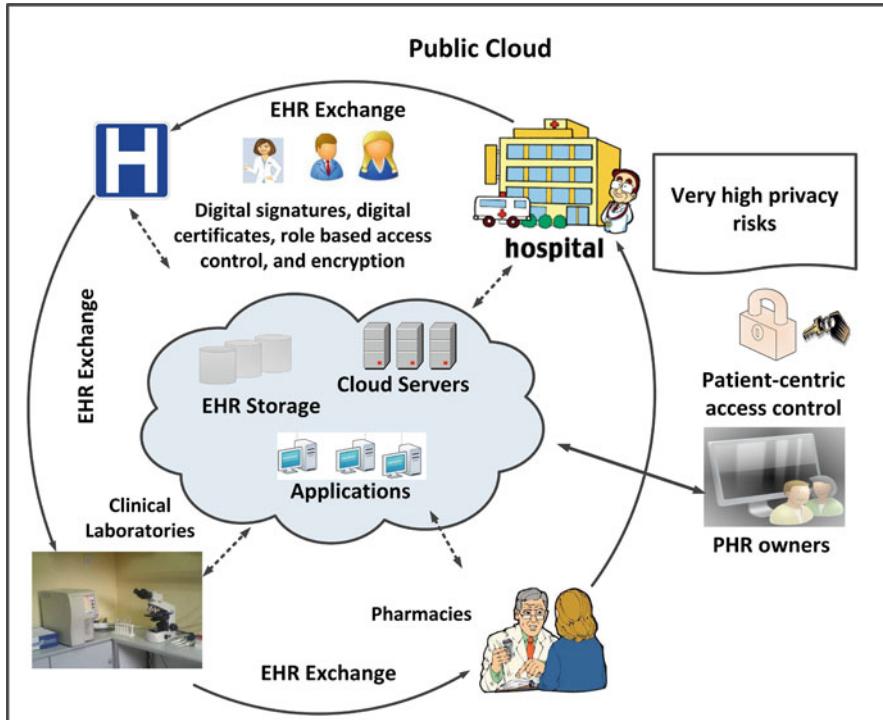


Fig. 15.3 An illustration of a public cloud in context of e-Health

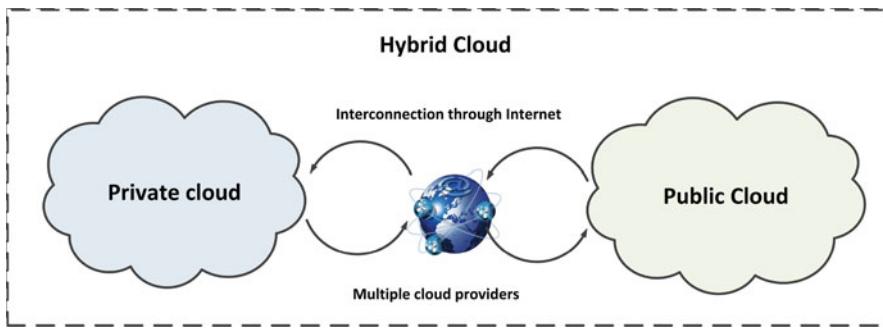


Fig. 15.4 An illustration of a hybrid cloud in context of e-Health

practitioner to a remote medical specialist [1]. Likewise, threats to health data may be internal [34]. For example, the CSP might learn the content of sensitive health data. Therefore, the enforced access control mechanisms must ensure that access to sensitive information should only be granted to entities having a right-to-know privilege.

15.2.3.1 Spoofing Identity

The spoofing identity threats include the unlawful attempts by other users, or machines, to pose as the valid users or machines [34]. In the e-Health cloud, the unlawful entities may obtain access to the patients' health data by spoofing the identities of the authorized users, such as doctors and patients themselves. Therefore, to counter the spoofing identity attacks, strong authentication mechanisms are necessary to restrict the unlawful data access.

15.2.3.2 Data Tampering

The malicious attempt to modify the data contents is called *data tampering* [67]. The health data is more vulnerable to tampering by both insiders and outsiders. The insiders include the hospital staff, pharmacy personnel, and insurance companies' staff that can modify the health data contents to obtain certain benefits. Moreover, the data is also vulnerable to tampering by other insiders, for example employees of the CSP who may act maliciously. Therefore, to ensure that the data present in the repositories is not modified through illegal means, data audit and robust accountability mechanisms are needed.

15.2.3.3 Repudiation

Repudiation refers to denying the obligations of a contract. The entities in a health cloud can falsely deny about the occurrence of an event with the health data contents [12]. Therefore, the data must be digitally signed to allow for maintaining evidence for all data manipulations.

15.2.3.4 Denial of Service (DoS)

In DoS attacks, the services are denied to privileged users [32]. Such issues arise due to the flaws in identity management schemes and the incompetence of the authorization mechanisms. Consequently, the legitimate users, for example the physicians or emergency staff, may not be able to access the cloud services when required.

15.2.3.5 Unlawful Privilege Escalation

Unlawful users may obtain access to data and can subsequently infiltrate into the system, such that the data contents at a large-scale are compromised [34]. The aforementioned data breaches in the healthcare domain are extremely critical and require the mechanisms to maintain the integrity of the health data.

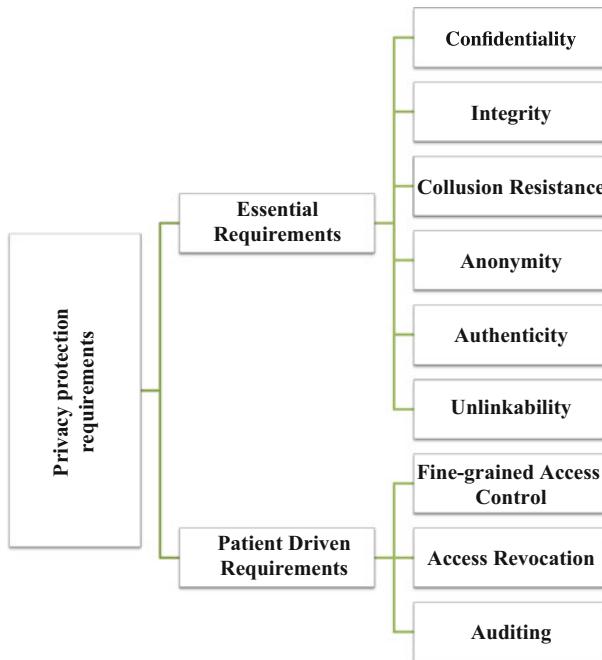


Fig. 15.5 Taxonomy of essential privacy requirements and patient-driven requirements

15.2.4 Essential Requirements for Privacy Protection

A model, called the CIA (confidentiality, integrity, availability)-triad, defines the necessary controls to maintain the privacy of data within an organization [36]. Data that are outsourced to third-party vendors, however, require more privacy measures than those specified in CIA. To deal with the threats presented in Sect. 15.2.3, the following privacy protection requirements should be fulfilled: (a) confidentiality, (b) integrity, (c) collusion resistance, (d) anonymity, (e) authenticity, and (f) unlinkability. Figure 15.5 presents a taxonomy of the privacy-protection requirements.

15.2.4.1 Confidentiality

The confidentiality requirement for the health data in the cloud environment requires that the data must be protected not only from the external entities, such as the CSP, but also from the unauthorized insiders [1].

15.2.4.2 Integrity

The integrity of the health data stored in the cloud requires the assurance that the data has not been modified through any illegitimate actions of either authorized or unauthorized users. In other words, the data present in the healthcare repositories must be a true characterization of the intended contents of the data.

15.2.4.3 Collusion Resistance

Collusion, in the context of cloud-based healthcare systems, refers to a mutual cooperation to learn the user-identities and contents of the health data illegitimately among the authorized and/or unauthorized entities. Therefore, collusion resistant approaches are inevitable to ensure the privacy of health data, not only from the external adversaries, but also from the valid insiders.

15.2.4.4 Anonymity

Anonymity refers to storing the health data contents in such a way that the identities (i.e., any unique identifiers, such as the name and social security number of a patient, or any other unique patient's information) of the subjects are properly concealed [14]. The patients' identities have to be protected from the CSPs, researchers, unprivileged users, and other malicious internal or external adversaries.

The methodologies used to maintain anonymity include the use of pseudonyms. Pseudonyms are identifiers that are used for identification of a subject instead of the subject's real name. It is also important to distinguish between the anonymity and pseudonymity, where pseudonymity is one of the methods currently used to offer anonymity. Moreover, anonymity requires that adversaries should not be able to infer meaningful information from the health data that would allow revealing data owners' identities.

15.2.4.5 Authenticity

Authenticity is necessary to ensure that access to health information is being requested by authorized users only. Only the entities possessing valid authentication codes and keys should be granted access to the health information.

15.2.4.6 Unlinkability

Unauthorized entities should not be able to infer any relationships between the (a) identifying information of the patients, such as the name and the address, and the (b) health information, for example diagnoses and medical histories. The adversaries

typically aim to observe the query patterns for certain records and attempt to interlink the patients' personal information with the patients' medical histories.

15.2.5 User/Patient Driven Privacy Protection Requirements

Besides the aforementioned privacy protection requirements, there are also certain patient-driven requirements that should be fulfilled to ensure fine-grained access control over the health data. These requirements include: (a) patient-centric access control, (b) access revocation, and (c) auditing. In what follows, we briefly elaborate on these requirements.

15.2.5.1 Patient-Centric Access Control

The users or patients can grant access to different entities over their health information, according to different access policies. The users can encrypt their PHRs before storing them in the cloud and access to this data can be permitted on a role basis. A Patient Controlled Encryption (PCE) based approach to delegate access rights to users has been presented in [10].

15.2.5.2 Access Revocation

The data owners or patients must be able to revoke the access rights granted to different entities over the health information in such a way that the users should not be able to access the health information anymore.

15.2.5.3 Auditing

The auditing of health data ensures that all activities and data manipulations of patients' health data are being monitored, either by the patients themselves or by any other entity nominated by the patients [44]. A similar concept to auditing is the *accountability* of health data, where data users are held responsible in the light of agreed-upon conditions [1].

15.2.6 Adversarial Models in the e-Health Cloud

The key participants of the cloud-based e-Health ecosystem include patients, health-care providers, pharmacies, clinical laboratories, and health insurance companies. The health data is shared among all of the aforementioned participating entities of

the cloud and, therefore, is vulnerable to theft and unauthorized disclosure [2]. The proposals developed to preserve privacy or mitigate security risks assume one of the following adversarial models: (a) trusted model, (b) untrusted model, and (c) semi-trusted model.

In the trusted adversarial model, the data stored on the cloud is mostly protected from external threats. Certain insiders, however, may behave maliciously and disclose sensitive information to unauthorized entities [50]. On the other hand, the untrusted adversarial model assumes threats to data privacy from both internal and external adversaries. As a result, the health data outsourced to the untrusted cloud storage requires strong security guarantees in order to fulfil the privacy requirements. In the semi-trusted adversarial model, the cloud servers are assumed to be honest in that they follow the protocol in general, but are curious to learn about the contents of the data [13, 58]. In order to mitigate the privacy disclosure risks in such models, identity management solutions that anonymize the identities of the patients and maintain the unlinkability of health information flows, are necessary.

15.3 Privacy Protection Strategies Employed in e-Health Cloud

To mitigate the privacy concerns related to the use of health data, both cryptographic and non-cryptographic approaches have been employed. The cryptographic approaches mostly utilize the Public Key Encryption (PKE), the Symmetric Key Encryption (SKE), and the El-Gamal cryptosystem. Other cryptographic primitives, such as Attribute Based Encryption (ABE) and its variants, Identity Based Encryption (IBE), Hierarchical Predicate Encryption (HPE), Proxy Re-encryption (PRE), and Homomorphic encryption have also been used. Similarly, to allow for user authentication, digital signatures and digital certificates have been employed. Interested readers are encouraged to consult [1] for a detailed presentation of each of the aforementioned cryptographic approaches. On the other hand, the non-cryptographic approaches use policy-based and broker-based authorization mechanisms to offer data privacy. In what follows, we discussed the approaches that are used to fulfil each of the essential requirements defined in Sect. 15.2.5.

15.3.1 Approaches to Protect Confidentiality in the e-Health Cloud

Various mechanisms have been employed to protect the confidentiality of health data in the cloud computing environment. Some recent approaches along with their strengths and weaknesses are presented below.

A platform for remote monitoring and secure exchange of health data in a cloud environment is proposed by Thilakanathan et al. [45]. The security protocol is implemented through an El-Gamal based proxy re-encryption methodology. The patients can transmit their health data to data consumers, such as doctors and nurses, in a secure manner. A key benefit of using proxy re-encryption is that the cipher-text generated under the public key of patient or data owner is translated by a semi-trusted proxy into a cipher-text that can be decrypted only by another user's private key. The construction does not permit any of the users to know the full decryption key. A trusted service, called Data Sharing Service (DSS), decrypts the keys using all of the pieces of the key corresponding to the requesting user. The decrypted keys are subsequently used to decrypt the data files shared in the cloud. The methodology also ensures revocation by simply removing the key pieces corresponding to the revoked users in the proxies, thereby eliminating the need to redistribute the keys to all of the users.

Han et al. [18] proposed a scheme to maintain the confidentiality of the health data in cloud-assisted Wireless Body Area Networks (WBANs). The scheme is intended to achieve secure communication between the WBANs and the cloud. The data that is encrypted by the users can be only decrypted by the intended recipients who possess the appropriate keys. The scheme utilizes multi-valued encoding rules that are equipped with ambiguous properties to circumvent attacks. The decoding process utilizes a complete finite graph to represent the encoded words correctly and utilizes the Dijkstra's algorithm [4] to determine a factorization that has high time complexity. Therefore, the approach can lead to additional overheads for encoding and decoding the exchanged messages.

Another approach to ensure the confidentiality of health data that is outsourced to a cloud environment is proposed by Tong et al. [47]. The proposed approach allows efficient searching on the encrypted data by using the Searchable Symmetric Encryption (SSE). To avoid the key wear-out, the approach frequently updates the keys. To enforce the auditability of the health data, the authors combined the threshold control signature with the ABE. Instead of delegating access control to individuals, a role-based access control approach is introduced to grant access in emergency situations where the trusted authorities can verify the signatures. Additionally, to ensure unlinkability of parts of health information, the key management is performed through a pseudorandom generator. Furthermore, the presented scheme offers search pattern privacy, anonymity, and keyword privacy. On the downside, the approach may be deficient in terms of dynamic access policy specification to grant role-based access due to its complex access structure.

An approach that delegates access control over the EHRs in public clouds, called Privacy-Preserving Attribute Based Group Key Management (PP AB-GKM), is presented in [35]. The authors employ a Two Layer Encryption (TLE) methodology, where the data owner performs a coarse-grained encryption while the fine-grained encryption is implemented by the cloud. A broadcast encryption methodology is used to transmit a message, after it is encrypted, to a subset of users. The methodology uses the Oblivious Commitment Based Envelop (OCBE) protocol to obliviously distribute the messages to users satisfying the particular access

conditions. Access Control Policies (ACPs) are enforced through attributes, such as role, insurance plan, type, and years of service. The policies for the role attribute are regulated by the data owners, whereas the cloud re-encrypts the data that is already encrypted by the data owner. The methodology effectively minimizes the overheads at the data owners' end, by reducing the number of attributes to be handled. In addition, the newly inserted group members, or revoked users, are effectively managed by the proposed methodology. On negative side, the proposed approach can result in key escrow problems at the CSP because of delegating partial encryption tasks to the cloud. Key escrow refers to the arrangement where the decryption keys are made available to certain other users, or third-parties, in addition to senders and receivers. In that sense, there are possibilities that the third-parties having access to the keys may misuse them.

In addition to the approaches discussed above, there are also methodologies to preserve the confidentiality of the health data in the cloud. For example, the authors in [37, 46] used the PKE approach, whereas [29, 64] used the SKE to ensure confidentiality. Table 15.1 presents a quick overview of the strategies that are used to maintain confidentiality of health data in cloud environments.

15.3.2 Approaches to Maintain Data Integrity in the e-Health Cloud

Various mechanisms have been employed to ensure the integrity of health data in the cloud computing environment. Some recent approaches, along with their strengths and weaknesses, are presented below.

A hybrid approach to preserving the privacy of health data that is shared in the cloud, has been proposed by Yang et al. [57]. The proposed model uses cryptography and statistical analysis to offer multi-level privacy. The medical datasets are partitioned vertically, such that on each partition of the EMR a different security level is implemented. The identifying attributes of the EMR, such as name, date of birth, and address, are encrypted using symmetric encryption. On the other hand, the portion of the EMR that consists of the clinical data and treatments' history is stored as plain text. Because the data is partitioned, it is difficult for the adversary to link the information. Moreover, only the recipients with appropriate authentication can merge the partitions through the decryption keys and quasi-identifiers. The data owners and the data recipients, respectively, ensure the integrity of the medical data locally and remotely. A limitation of the proposed approach is that the data recipient, who in this case is a cloud provider, can act maliciously and disclose the information that can help linking the portions of patients' medical records.

A secure and scalable cloud-based architecture for medical wireless networks has been proposed by Lounis et al. [32]. The proposed approach offers integrity of the outsourced medical data and a fine-grained access control is implemented through

Table 15.1 Overview of the approaches employed to maintain confidentiality

References	Encryption type	Strength(s)	Weakness(es)	TM	AM	PAC	AT	RV	Others
[45]	Proxy re-encryption, El-Gamal cryptosystem	Easy revocation	Less support for complex access policies	T	Password/PIN	–	–	✓	CR
[18]	SKE	Ensures data confidentiality	High encryption and decryption overheads	T	–	–	–	–	–
[47]	ABE	Keyword search privacy	Dynamic access policies	T	Digital signature	–	–	✓	UN, AN
[35]	PKE	Dynamic policies	Key escrow issues	U	–	✓	–	✓	–

Symbol “✓” represents that a particular patient-driven requirement is fulfilled by the technique; symbol “–” represents that a particular feature/requirement is not fulfilled, or is not addressed by the authors of the technique. The following abbreviations are used: threat model *TM*, trusted *T*, untrusted *U*, semi-trusted *S*, authentication mechanism *AM*, patient-centric access control *PAC*, audit *AT*, revocation *RV*, confidentiality *CO*, integrity *IN*, collusion resistance *CR*, anonymity *AN*, unlinkability *UN*

the CP-ABE based construction. The system utilizes the Wireless Sensor Network (WSN) to collect data from the patients, whereas the data stored on cloud servers are accessed by the healthcare professionals through an application. To support access policies in the medical domain, where certain users are granted access while others are not, the authors used ABE and symmetric cryptography. The data file, before storage in the cloud, is encrypted through a symmetric key and the keys are further encrypted through ABE. Only the users possessing the secret keys satisfying the access policies can decrypt the data files. In addition, the authors introduced a role, called Healthcare Authority (HA), to define and enforce the policies and to generate the security parameters. A key benefit of this approach is that in the case of change of access policies, it requires the re-encryption of only the keys, which significantly results in a reduction of the encryption overheads of ABE. However, the approach can come across the management issues arising due to the frequent policy changes, particularly in the case of access revocation.

A proposal to enhance the integrity and accountability of the EHRs by enforcing either the explicit or implicit patient control over the EHRs, is presented by Mashima et al. [33]. A monitoring agent monitors how and when the data is accessed. The authors employed PKE to encrypt the health records. To make sure that only the valid and trusted entities access and use the health records, the scheme employs a standard digital signature approach, called Designated Verifier Signatures (UDVS). In addition, besides encrypting and digital signing, the entities interested in accessing the health records must contact the monitoring agent to confirm that they intend to use the records. On the other hand, the authors' assumption that the record issuers are knowledgeable about the record contents and keys, can result in information disclosure as a result of any malicious activity by the issuers.

Wang et al. [51] have presented a scheme to circumvent tampering attempts to the health data outsourced to third-party cloud servers. The proposed scheme deploys an independent third-party to maintain health data integrity and prevents illegitimate data modification attempts by the cloud providers, hospitals, and patients. Before being uploaded to the cloud, the data is encrypted by the patients using SKE under the private keys of the patients. Homomorphic verifiable tags are used to ensure the computations made on the encrypted data. In addition, a Diffie-Hellman based key exchange strategy is employed to securely exchange the keys in a multi-member scenario. Nonetheless, the large cipher-texts that are particular for homomorphic encryption can result in performance issues. The authors in [30] also used homomorphic encryption with IBE to preserve patients' privacy in the context of a mobile health monitoring system.

In addition to the above-mentioned approaches, few other strategies have also been proposed to maintain the integrity of health data in a cloud environment. The authors in [23] used PKE and digital signatures, whereas Hierarchical Predicate Encryption (HPE) and ABE have been used in [7]. Similarly, a policy-based authorization methodology has been proposed in [17] to help maintain data integrity. Table 15.2 presents a summary of the approaches fulfilling the integrity requirement.

Table 15.2 Overview of the approaches employed to maintain data integrity

References	Encryption type	Strength(s)	Weakness(es)	TM	AM	PAC	AT	RV	Others
[57]	SKE	Offers multi-level privacy	Possible disclosure by CSP	S	–	✓	–	✓	–
[32]	ABE, SKE	Fine-grained access control	Policy specification issues	U	Digital certificate	✓	–	✓	CO
[33]	PKE	Monitoring of data while being used	Record issuer may disclose sensitive information	U	Digital signature	–	✓	–	AU
[51]	PKE	Effective against tampering	Increased key management overheads	U	Digital signatures	–	–	–	–

Symbol “✓” represents that a particular patient-driven requirement is fulfilled by the technique; symbol “–” represents that a particular feature/requirement is not fulfilled, or is not addressed by the authors of the technique. The following abbreviations are used: threat model *TM*, trusted *T*, untrusted *U*, semi-trusted *S*, authentication mechanism *AM*, patient-centric access control *PAC*, audit *AT*, revocation *RV*, confidentiality *CO*, integrity *IN*, collusion resistance *CR*, anonymity *AN*, unlinkability *UN*

15.3.3 *Approaches to Offer Collusion Resistance in the e-Health Cloud*

Numerous approaches have been proposed to ensure collusion resistance in an e-Health cloud. Some recent approaches that were developed to deal with the collusion attacks are presented below.

A scheme to resist collusion attacks by the semi-trusted CSP was proposed in [13]. This approach enforces access control on the attributes of the health data files that are stored in the could, through the use of CP-ABE and IBE. A public-private key pair for each attribute is defined and the secret key of a user is constructed such that it is a combination of the public key of the user and the secret key of the attribute. The users can decrypt only those cipher-texts that match with the defined access structure. The scheme is collusion resistant in a way that even if the unauthorized users collude with the authorized users, it is impossible for them to get a clue about the contents of the data file. The resistance against collusion is enforced by associating the public keys of the users with the respective attributes to stop combining with the other users' attributes. Although the approach claims to avoid key management issues by offering key assignment through the cloud, key escrow problems may exist due to the assumption of a semi-trusted adversarial model.

A methodology, called Priority based Health Data Aggregation (PHDA), that preserves the identity and data privacy in transmissions from the WBANs, is presented in [63]. The proposed scheme employs the Paillier cryptosystem to resist against eavesdropping. To gain access to the health data of a particular patient, the requesting entity submits a request to the mobile users through the CSP. Only the entities possessing the valid secret keys are granted access to the health data. Moreover, the identities of the patients are anonymized by changing the pseudonyms in different periods of time and only the trusted authority has the ability to link different pseudonyms. Furthermore, the methodology offers protection against forgery attempts by malicious insiders, as the trusted authority authenticates the correctness of the data requested by the doctors before transmission. However, a limitation of the methodology is that it incurs significant overheads in terms of key management.

Wang et al. [52] proposed an attribute-based method to enforce patient-centric access control to health data stored in the cloud, through constant-size Ciphertext Policy Comparative Attribute-Based Encryption (CCP-CABE). The framework introduces a trusted authority that is responsible for generating the encryption and decryption keys, and associated parameters. Another trusted authority, called the encryption service provider, is used to help data owners generate partial encrypted headers based on the attributes specified in the access policies. The proposed method is claimed to be effective for collusion attacks by honest-but-curious cloud servers. A key benefit of the approach is that it attempts to minimize the encryption overhead at the data owner side, where the data owners can partially delegate the encryption tasks to the encryption service providers. In addition, fixed sizes of ciphertexts are

useful in reducing the overheads that are exist because of the CP-ABE. On the negative side, the approach permits other roles, for example nurses, to delegate access to health data, which may lead to sensitive information disclosure.

Another CP-ABE based approach, called the Efficient and Secure Patient-centric Access control Scheme (ESPAC) for cloud storage, was proposed in [9]. This approach implements patient-centric access control and utilizes IBE to establish a secure communication between the remote patients and the CSP. To establish the access structure, the patients transmit the secret keys to the CSP. The approach ensures integrity of the transmitted message, authenticity of the message originator, non-repudiation, and is also resistant to collusion and Denial of Service (DOS) attacks. The performance results show that the ESPAC scheme can resist DOS attacks in a dual server mode. The lack of dynamicity and flexibility in patient data attainment and transmission to the hospital servers, constitutes this approach inefficient. Moreover, the hospital servers can be a bottleneck of the system, and in the event of failure the access to the data stored on the cloud may be restricted.

Liu et al. [31] proposed a cloud based patient-centric Clinical Decision Support System (CDSS) and claimed that the privacy of patients' data obtained during the clinical visits is preserved. In this approach, diagnosis is performed by mapping the symptom of the patients to past patients through a Naïve Bayes classifier. The data of the existing patients is stored on the cloud. However, storing such huge volumes of private data on third-party servers entails serious threats of data theft and collusion. The approach proposed in [31] offers privacy for the historical data of existing patients, by employing an El-Gamal encryption based methodology, called Additive Homomorphic Proxy Aggregation (AHPA). The model mainly comprises of entities, such as Trusted Authority (TA), Cloud Platform (CP), Data Provider (DP), and Processing Unit (PU). The approach assumes an honest-but-curious server that generally follows the protocol. It effectively deals with collusion attacks both by the CP and the PU, by performing a re-encryption of the ciphertext, which makes decryption impossible for the adversary without knowing the private key. However, a limitation of this approach is that it may incur high communication overheads because of the key generation and re-encryption operations. Table 15.3 presents a summary of the approaches developed to overcome the collusion.

15.3.4 Approaches to Maintain Anonymity in the e-Health Cloud

Various mechanisms have been proposed to maintain the anonymity of health data in the cloud environment. Some recently developed approaches, along with their strengths and weaknesses, are presented below.

Shen et al. [40] proposed a health monitoring architecture for geo-distributed clouds. To circumvent the identification of patient identities through traffic analysis attacks, the authors employed a traffic-shaping algorithm. The shaping algorithm equally distributes both the health data traffic and the non-health data traffic.

Table 15.3 Overview of the approaches employed to offer collusion resistance

References	Encryption type	Strength(s)	Weakness(es)	TM	AM	PAC	AT	RV	Others
[13]	CP-ABE	Minimal key management overheads	Key escrow	S	—	—	—	—	—
[63]	Paillier cryptograph, PKE	Priority based data aggregation	Key management issues	S	Secret keys, certificates	—	✓	✓	—
[52]	ABE	Constant sized cipher-texts	Issues with delegation of access controls	S	—	—	—	✓	UN
[9]	CP-ABE	Patient centric access control	Hospital servers may be single point of failure	T	Digital signatures	✓	—	—	CO, AU

Symbol “✓” represents that a particular patient-driven requirement is fulfilled by the technique; symbol “—” represents that a particular feature/requirement is not fulfilled, or is not addressed by the authors of the technique. The following abbreviations are used: threat model *TM*, trusted *T*, untrusted *U*, semi-trusted *S*, authentication mechanism *AM*, patient-centric access control *PAC*, audit *AT*, revocation *RV*, confidentiality *CO*, integrity *IN*, collusion resistance *CR*, anonymity *AN*, authenticity *AU*, unlinkability *UN*.

The autocorrelations obtained for the shaped health data traffic are observed close to the non-health data traffic. The authors analyzed the capability of their traffic analyzer by employing the Kullback-Leibler (K-L) divergence. K-L divergence is an entropy measure used to quantify the difference between two probability distributions [27]. The approach of Shen et al. [40] is claimed to be effective for privacy protection to a reasonable degree. A limitation of the approach is that the shaping algorithm incurs high communication delays because of the geo-distributed clouds.

Zhang et al. [65] proposed a secure method to form a social group of patients suffering from similar diseases, in cloud-assisted WBANs. The proposed approach divides the whole district into different blocks, where each block has a designated block manager and patients with mobile devices (sinks) move around the blocks. The block managers sense and collect the Personal Health Information (PHI), such as the Electrocardiogram (ECG) and Electroencephalography (EEG) data from the mobile devices of the patients. The block managers also update and distribute the private keys. The considered threat model assumes that adversaries may compromise more sensors than the threshold in a fixed block location and in a single period of time from different blocks. The methodology proposes the use of resource constrained mobile devices that are less vulnerable to attacks and considers that the block managers are trusted entities. A secure communication channel is established through the Diffie-Hellman key exchange protocol among the sink nodes and the block managers. To maintain the anonymity of the on-body sensors and the patients, identity blinding matrices are employed. In addition, the compromised on-body sensors in the event of a compromise are revoked and cannot be further exploited to derive the keys. A shortcoming of the proposed approach is that it does not offer any solution for valid insiders who may behave in a malicious way and help adversaries by disclosing the sensitive health information or the keys.

The authors of [14] designed a methodology for secure sharing of data in multi-cloud environments. The proposed approach enforces role-based access control through Ciphertext Policy Attribute Based Encryption (CP-ABE). A key benefit of CP-ABE in the proposed multi-cloud scenario is that only the delegated users can have access to specified attributes. The patients' health data are signed and encrypted by the doctor, and subsequently sent to a local Multi-cloud Proxy (MCP). The MCP splits the health record according to a secret sharing scheme and each portion of the record is subsequently stored at different cloud providers' locations. Splitting the health record into multiple portions not only enhances the unlinkability of the health data and patients' personal information, but also has a positive effect on patients' anonymity. Moreover, to ensure that the privacy of portions stored at different cloud providers is not compromised, the approach employs hash-based identifiers. The cloud provider, or any adversary, is not able to deduce the meaningful information from the hash because cryptographic hashes are difficult to invert. However, the approach is limited in user revocation and in handling emergency expectations. In addition, the complexity of searching for the identifiers from different cloud providers while reconstructing the health records, also results in an increased computational overhead.

Xu et al. [56] propose an pseudonymization approach for the secondary use of EHRs stored in a cloud infrastructure. In this paper, the authors propose to remove all third-parties that are potentially considered as malicious. A trusted authority issues the certificates to all the entities, such as doctors, pharmacists, and insurance companies. Each of the requesting user/entity generates his/her private key that subsequently, along with the user name and other information, is digitally signed by the trusted authority. On the other hand, the patients generate the pairs of private and public keys. The public keys are made available to the insurance companies to get the certificate, whereas the private key, which is only known to the patient, is stored at the protected memory of the patient's smart card. However, the certificate issued by the insurance company does not contain patient's identifying information. As a result, it can be difficult for the scheme to differentiate between the valid certificate applicants and malicious insiders, such as employees of the insurance company, that eventually can result in not only revealing the patients' identities but also disclosing primary health information.

Similar approaches for offering anonymity in a cloud environment were proposed in [39], where the authors used ABE, and in [16], where the SSH protocol was used for this purpose. Table 15.4 presents a summary of the approaches developed to maintain anonymity in a health cloud environment.

15.3.5 Approaches to Offer Authenticity in the e-Health Cloud

Several methods have been proposed to ensure the authenticity of health data in a cloud environment. Some recently developed approaches, along with their strengths and weaknesses, are discussed below.

The authors of [43] propose a standard model for the automated collection of health data, generated by personal devices, such as wireless glucometers and smartphones of the patients. The collected data is transmitted to the authorized providers. The proposed model allows the patients to exercise access control on their health data. The data requested by the clinicians are encrypted, as well as the S/MIME attachments. The sending entities sign the messages to ensure that the messages have not been tampered during transit. The tasks of message encryption and digital signing are accomplished through robust public key infrastructure-based protocols. Each message contains a certificate created by a digital certificate authority. Nevertheless, the DIRECT messaging to access request does not specify whether a certificate authority is trusted or not and, therefore, can potentially lead to tampering.

Yu et al. [59] presented a methodology that permits the physicians to collect the patients' evidence-based data and share it with other physicians. The privacy of the data is preserved through a secure authentication mechanism. The new physicians in the system are authenticated through the Health Personal Cards (HPCs). For the existing physicians, the access to the health data is provided through the Secure Sockets Layer (SSL) protocol. SSL establishes encrypted communication

Table 15.4 Overview of the approaches employed to maintain anonymity

References	Encryption type	Strength(s)	Weakness(es)	TM	AM	PAC	AT	RV	Others
[40]	Not used	Privacy preservation without encryption	High communication delays	U	–	–	–	–	–
[65]	SKE	Efficient key management	Lacks means to deal with malicious insiders	T	Digital signature	–	–	✓	CO
[14]	CP-ABE	Secure sharing in hybrid cloud	Inefficient revocation	S	Digital signature	✓	–	–	–
[56]	PKE	Easy pseudonym generation	Difficult to differentiate between valid and malicious users during certificate generation	S	Digital certificate	–	–	–	–

Symbol “✓” represents that a particular patient-driven requirement is fulfilled by the technique; symbol “–” represents that a particular feature/requirement is not fulfilled, or is not addressed by the authors of the technique. The following abbreviations are used: threat model *TM*, trusted *T*, untrusted *U*, semi-trusted *S*, authentication mechanism *AM*, patient-centric access control *PAC*, audit *AT*, revocation *RV*, confidentiality *CO*, integrity *IN*, collusion resistance *CR*, anonymity *AN*, authenticity *AU*, unlinkability *UN*.

between the servers and the browser. In addition, patients' identifying information is encrypted through the Advanced Encryption Standard (AES) of the symmetric key encryption, which helps in maintaining unlinkability.

Zhou et al. [66] proposed a scheme that aims to offer confidentiality and anonymity to patients in an m-healthcare cloud system. An Attribute Based Designated Verifier Signature scheme (ADVS) is proposed to ensure the patient-self controllable access over the health information. The patients encrypt the PHI, instead of assigning the secret keys to each of the physicians; define the access policies for different user groups or classes. Access control is realized for the following groups: (a) directly authorized physicians, (b) indirectly authorized physicians, and (c) unauthorized persons or adversaries. The authorized physicians are provided complete access to the patients' health information and are also aware of the identities of the patients. Although the indirectly authorized physicians can view the health information, they are unable to learn the identities of the patients. On the other hand, the unauthorized patients cannot have access to any information. The users from each group need to satisfy the access structure defined on the respective attributes. The scheme permits the patients to generate the attribute-based signatures for each group that is different from the others and are not linkable in any way. A key strength of the approach is that despite of using the attribute-based access structure, the computational and communication overheads are low. On the negative side, it is difficult to revoke access rights from the indirectly authorized physicians.

The authors of [21] used Public Key Encryption (PKE) and Content Key Encryption (CKE) to fulfil the authenticity requirement. Yu et al. [58] used Key Policy-Attribute Based Encryption (KP-ABE), proxy re-encryption, and lazy re-encryption to achieve secure and scalable access control and to ensure the authenticity of the health information. Table 15.5 presents a summary of the approaches developed to maintain authenticity of health data.

15.3.6 Approaches to Maintain Unlinkability in the e-Health Cloud

Maintaining unlinkability between the patients' medical records, such as diagnosis history, treatments, and lab reports, and the patients' identities, is an important requirement to be fulfilled in order to protect patients' privacy. Some recently developed approaches, along with their strengths and weaknesses, are presented below. Table 15.6 presents a summary of the approaches developed to maintain the unlinkability between the health data and patients' identifying information.

A semantic privacy management framework for facilitating the secure sharing of health data through mobile applications is presented in [8]. The framework utilizes the XACML privacy policy language to enforce a set of access policies. The approach is claimed to be effective in identifying the attacks by adversaries asking "similar" health data in different queries. Each query by the adversary is

Table 15.5 Overview of the approaches employed to maintain authenticity

References	Encryption type	Strength(s)	Weakness(es)	TM	AM	PAC	AT	RV	Others
[43]	PKE	Introduces a standard for interoperability between the devices and the EHRs	Vulnerable to tampering	T	Digital signature	✓	–	–	–
[59]	SKE	Efficient against eavesdropping	Interoperability issues with hospital information systems	–	Digital certificate	–	–	–	–
[66]	PKE	Multi-level privacy protection	Computational overhead	U	Semi-trusted	✓	–	–	CO, AN

Symbol “✓” represents that a particular patient-driven requirement is fulfilled by the technique; symbol “–” represents that a particular feature/requirement is not fulfilled, or is not addressed by the authors of the technique. The following abbreviations are used: threat model *TM*, trusted *T*, untrusted *U*, semi-trusted *S*, authentication mechanism *AM*, patient-centric access control *PAC*, audit *AT*, revocation *RV*, confidentiality *CO*, integrity *IN*, collusion resistance *CR*, anonymity *AN*, authenticity *AU*, unlinkability *UN*.

Table 15.6 Overview of the approaches employed to maintain unlinkability

References	Encryption type	Strength(s)	Weakness(es)	TM	AM	PAC	AT	RV	Others
[8]	Not used	Semantics driven policy enforcement	Integrity and authenticity of stored data	T	–	✓	–	–	AN
[5]	PKE	Selective access control	Key revocation and integrity issues	S	–	–	–	–	AN
[55]	PKE	Identity privacy	Difficulty in management of multiple operations	U	Smart card and PIN	–	–	–	AN
[11]	Searchable encryption, PKE	Searching over encrypted data	Lack of access control	S	Not indicated	–	–	–	AN

Symbol “✓” represents that a particular patient-driven requirement is fulfilled by the technique; symbol “–” represents that a particular feature/requirement is not fulfilled, or is not addressed by the authors of the technique. The following abbreviations are used: threat model *TM*, trusted *T*, untrusted *U*, semi-trusted *S*, authentication mechanism *AM*, patient-centric access control *PAC*, audit *AT*, revocation *RV*, confidentiality *CO*, integrity *IN*, collusion resistance *CR*, anonymity *AN*, authenticity *AU*, unlinkability *UN*.

assumed to have some changed requested parameters. The users' data is stored in the data store and the access policies are associated with the data. The users request access to data via their mobile devices. Each request is forwarded to the Policy Enforcement Point (PEP) that further forwards it to the Policy Decision Point (PDP). The policies associated with the requested data are evaluated by the semantic handler and the context handler to classify whether the query is malicious or legitimate. The approach enforces access policies to ensure fine-grained access control, but may possibly come across integrity and authenticity issues for the data stored in the cloud.

Ahmad et al. [5] proposed a methodology that aims to minimize the cost of encryption, decryption, and key management of health data in the cloud. The methodology, called Dual Lock, allows access to PHI by applying dual keys to offer protection against attacks issued by honest-but-curious cloud servers. The methodology is also secure against malicious attacks that attempt to gain information through the traffic patterns. For each query, a random output pattern is produced that stops the adversary from inferring and linking the information.

A decentralized pseudonym scheme to protect the patients' identities while storing the PHRs in the cloud, is proposed by Xu et al. [55]. The patients in the proposed model visit the doctors with their smart cards. The doctors after examining the patients they may need to prescribe medicine to be purchased from the pharmacy. The pseudonym and the encryption key are generated by using the patients' smart card. The doctor encrypts and signs the prescription record and the diagnosis data, and the data is uploaded to the cloud. The cloud, with the help of a trusted authority, before storing the EHRs also verifies the certificates of the doctors and challenges the doctors' private keys. The strategy restricts entities other than the patients to generate valid encryption keys and pseudonyms. The patients can easily identify and remove any forged EHR by inspecting the pseudonym and the encryption key. The cloud has no knowledge of the contents of the EHRs, as the important contents are encrypted and the keys are unknown to the cloud. The scheme also involves the patients and doctors to construct the search indexes to support complex searches on the encrypted data in the cloud. Moreover, the approach allows the patients to frequently change their pseudonyms to avoid identify theft. On negative side, the scheme requires complex operations for encryption key generation, certificate creation and verification, and pseudonym generation, that seem very difficult to manage in a real-time setting.

Cao et al. [11] proposed a scheme, called Multi-keyword Ranked Search over Encrypted data (MRSE), to enable search operations while preserving privacy. The scheme is applicable to the PHRs and can restrict the entities searching the health data from learning the patients' identifying information. The scheme assumes an honest-but-curious cloud server that generally follows the protocol but is also interested in learning additional information. In the proposed scheme, the cloud servers only possess the information about the encrypted data and the searchable index of the outsourced data. The scheme restricts the semi-trusted cloud servers to generate trapdoors and to derive any information from the search keywords and queries. The authors utilized the inner product similarity computation that is

based on the secure k -nearest neighbour (k -NN) technique. A shortcoming of the proposed scheme is that it is limited in enforcing the user-centric access control. The approaches presented in [30, 37] also ensure the unlinkability between the patients' health information and personal identifying information.

15.4 Discussion and Open Research Issues

The techniques presented in Sect. 15.3 effectively counter the threats, such as spoofing identity, tampering, repudiation, denial of service attacks, and unlawful privilege escalation. In addition, the discussed techniques also sufficiently fulfil the privacy protection requirements that arise due to the architecture of the cloud. The PKE based approaches were observed to be amply satisfying the unlinkability, collusion, and anonymity requirements as compared to the SKE based constructions. However, considering the widespread use of the public key infrastructure in most of the Web based architectures, there is still a need to enhance the capabilities of PKE to ensure the confidentiality, integrity, and authenticity of the health-data in the public cloud environment.

Interestingly, the ABE-based schemes fulfilled most of the privacy protection requirements. In addition, the ABE approaches also offered reasonable patient-centric and fine-grained access control. However, such approaches generally come across the issue of management of access policies. Moreover, the ABE-based schemes tend to be expensive in terms of computational complexity, because they usually perform bitwise matching of attributes in a hierarchical structure. Consequently, the ciphertext size and key size increase linearly with the number of attributes. Furthermore, the cipher-texts in ABE are of larger size and also decryption operations are computationally expensive. This makes the use of ABE questionable in power and energy constrained handheld devices. El-Gamal cryptosystem and proxy re-encryption based approaches also proved very effective to fulfil the important privacy protection requirements. However, they also have the same issues as the ABE approaches.

Despite of the effectiveness of the methodologies and constructions employed to protect the privacy of the cloud based e-Health systems, there are certain issues that need more attention to help mitigate the privacy concerns.

The procedures to ensure the audits and accountability of health data utilization need to be strengthened. Enforcing the audit and monitoring on the use of health data can not only warrant the integrity and confidentiality of the health data but can also help identifying how the privileged entities are utilizing the healthcare data. The development of procedures to identify or distinguish whether an entity is following the protocol in the usual way or not, is still an open issue that has to be explored.

Another important issue worth exploring is revocation of access rights granted to the privileged entities over the health data. Currently, the majority of the proposals implementing the patient-centric access control, employ ABE. However, due to the complex nature of the e-Health cloud, the specification and implementation of the access policies for a diverse type of users is indeed a challenging task. The management and update of access key shares for the departing or revoked users should be done in such a way that the need to redistribute the keys is eliminated.

Although the homomorphic encryption allows computations over encrypted data, the computational complexity of such operations enormously grows and the approach seems impractical to handle a large number of operations on encrypted data. Indeed, it currently seems unrealistic to perform computations over encrypted data in real-life large-scale systems. Consequently, mechanisms are needed that ensure privacy while computations are being performed over unencrypted data.

Another significant direction for future research is to devise methodologies that first determine the importance of the data to be encrypted in the cloud environment, and then encrypt only those portions of the data that contain extremely sensitive information. The rationale for partially encrypting the data is the increased cost of applying the encryption process. A few related solutions [38, 61, 62] have already been developed. In any case, the practicality of such methodologies needs to be portrayed to allow for their extended adoption, because users and healthcare providers will not rely on solutions whose security strengths are not formally proven.

In conclusion, the e-Health cloud is an exciting opportunity for healthcare providers. However, the privacy of health data in cloud environments has emerged as the most important concern that needs particular attention in order to ensure the widespread acceptance of the cloud computing model in the healthcare domain.

15.5 Conclusion

The distributed architecture of the cloud poses serious questions on the privacy of sensitive health data stored in the cloud. This chapter presented the privacy threats to the electronic health data and discussed recent solutions to counter these threats. The analysis of the discussed strategies indicates that the cryptographic solutions, along with digital signatures and certificate-based authentication mechanisms, are capable enough to provide privacy guarantees in a public cloud environment. However, still robust methodologies are needed to ensure the secure provenance and monitoring of electronic health data. The e-Health clouds can gain worldwide acceptance only if they win the confidence and trust of healthcare organizations and patients, by providing strong mechanisms to protect the sensitive electronic health information.

References

1. Abbas, A., Khan, S.: A review on the state-of-the-art privacy preserving approaches in e-health clouds. *IEEE J. Biomed. Health Inform.* **18**, 1431–1441 (2014)
2. Abbas, A., Bilal, K., Zhang, L., Khan, S.U.: A cloud based health insurance plan recommendation system: a user centered approach. *Futur. Gener. Comput. Syst.* **43**, 99–109 (2015)
3. Abbas, A., Khan, M., Ali, M., Khan, S., Yang, L.: A cloud based framework for identification of influential health experts from Twitter. In: Proceedings of the 15th International Conference on Scalable Computing and Communications (ScalCom) (2015)
4. Sedgewick, R.: Algorithms in C++: Fundamentals, data structures, sorting, searching and graph algorithms. Boston: Addison-Wesley (2001)
5. Ahmad, M., Pervez, Z., Lee, S.: Dual locks: partial sharing of health documents in cloud. In: Bodine, C., Helal, S., Gu, T., Mokhtari, M. (eds.) Smart Homes and Health Telematics, vol. 8456, pp. 187–194. Springer, Cham (2014)
6. Ahuja, S., Mani, S., Zambrano, J.: A survey of the state of cloud computing in healthcare. *Netw. Commun. Technol.* **1**, 12–19 (2012)
7. Akinyele, J., Lehmann, C., Green, M., Pagano, M., Peterson, Z., Rubin, A.: Self-protecting electronic medical records using attribute-based encryption. Technical Report 2010/565, Cryptology e-Print Archive (2010)
8. Ammar, N., Malik, Z., Rezgui, A., Alodib, M.: MobiDyC: private mobile-based health data sharing through dynamic context handling. *Procedia Comput. Sci.* **34**, 426–433 (2014)
9. Barua, M., Liang, X., Lu, R., Shen, X.: ESPAC: enabling security and patient-centric access control for eHealth in cloud computing. *Int. J. Secur. Netw.* **6**(2/3), 67–76 (2011)
10. Benaloh, J., Chase, M., Horvitz, E., Lauter, K.: Patient controlled encryption: ensuring privacy of electronic medical records. In: Proceedings of the 2009 ACM Workshop on Cloud Computing Security, CCSW ’09, pp. 103–114. ACM (2009)
11. Cao, N., Wang, C., Li, M., Ren, K., Lou, W.: Privacy-preserving multi-keyword ranked search over encrypted cloud data. *IEEE Trans. Parallel Distrib. Syst.* **25**(1), 222–233 (2014)
12. Chen, Y., Lu, J., Jan, J.: A secure EHR system based on hybrid clouds. *J. Med. Syst.* **36**(5), 3375–3384 (2012)
13. Dong, X., Yu, J., Luo, Y., Chen, Y., Xue, G., Li, M.: Achieving an effective, scalable and privacy-preserving data sharing service in cloud computing. *Comput. Secur.* **42**, 151–164 (2014)
14. Fabian, B., Ermakova, T., Junghanns, P.: Collaborative and secure sharing of healthcare data in multi-clouds. *Inf. Syst.* **48**, 132–150 (2015)
15. Fan, L., Buchanan, W., Thummel, C., Lo, O., Khedim, A., Uthmani, O., Lawson, A., Bell, D.: DACAR platform for e-Health services cloud. In: Proceedings of the 4th IEEE International Conference on Cloud Computing, pp. 219–226 (2011)
16. Gorp, P., Comuzzi, M.: Lifelong personal health data and application software via virtual machines in the cloud. *IEEE J. Biomed. Health Inform.* **18**(1), 36–45 (2014)
17. Haas, S., Wohlgemuth, S., Echizen, I., Sonehara, N., Muller, G.: Aspects of privacy for electronic health records. *Int. J. Med. Inform.* **80**(2), e26–e31 (2011)
18. Han, N., Han, L., Tuan, D., In, H., Jo, M.: A scheme for data confidentiality in cloud-assisted wireless body area networks. *Inf. Sci.* **284**, 157–166 (2014)
19. Hans, L., Sadeghi, A., Winandy, M.: Securing the e-Health cloud. In: Proceedings of the 1st ACM International Health Informatics Symposium, IHI ’10, pp. 220–229. ACM (2010)
20. HealthIT.gov.: Health IT legislation and regulations. <http://healthit.gov/policy-researchers-implementers/health-it-legislation> (2015). Accessed 29 May 2015
21. Jafari, M., Naini, R., Sheppard, N.: A rights management approach to protection of privacy in a cloud of electronic health records. In: Proceedings of the 11th Annual ACM Workshop on Digital Rights Management, DRM ’11, pp. 23–30. ACM (2011)

22. Johnson, M.: Data hemorrhages in the health-care sector. In: Dingledine, R., Golle, P. (eds.) *Financial Cryptography and Data Security*. Lecture Notes in Computer Science, vol. 5628, pp. 71–89. Springer, Berlin/Heidelberg (2009)
23. Kaletsch, A., Sunyaev, A.: Privacy engineering: personal health records in cloud computing environments. In: Proceedings of the 32nd International Conference on Information Systems (ICIS), pp. 1–11 (2011)
24. Kamara, S., Lauter, K.: Cryptographic cloud storage. In: Sion, R., Curtmola, R., Dietrich, S., Kiayias, A., Miret, J., Sako, K., Kazue, S. (eds.) *Financial Cryptography and Data Security*. Lecture Notes in Computer Science, vol. 6054, pp. 136–149. Springer, Berlin/Heidelberg (2010)
25. Kuo, A.H.: Opportunities and challenges of cloud computing to improve health care services. *J. Med. Internet Res.* **13**(3), e67 (2011). doi:[10.2196/jmir.1867](https://doi.org/10.2196/jmir.1867)
26. Li, J.: Electronic personal health records and the question of privacy. *Computer* **99** (2013). doi:[10.1109/MC.2013.225](https://doi.org/10.1109/MC.2013.225)
27. Li, X.B., Sarkar, S.: Against classification attacks: a decision tree pruning approach to privacy protection in data mining. *Oper. Res.* **57**(6), 1496–1509 (2009)
28. Li, M., Yu, S., Ren, K., Lou, W.: Securing personal health records in cloud computing: patient-centric and fine-grained data access control in multi-owner settings. In: *Security and Privacy in Communication Networks*, Springer Berlin Heidelberg, pp. 89–106 (2010)
29. Li, Z., Chang, E., Huang, K., Lai, F.: A secure electronic medical record sharing mechanism in the cloud computing platform. In: Proceedings of the 15th IEEE International Symposium on Consumer Electronics (ISCE), pp. 98–103. ACM (2011)
30. Lin, H., Shao, J., Zhang, C., Fang, Y.: CAM: cloud-assisted privacy preserving mobile health monitoring. *IEEE Trans. Inf. Forensics Secur.* **8**(6), 985–997 (2013)
31. Liu, X., Lu, R., Ma, J., Chen, L., Qin, B.: Privacy-preserving patient-centric clinical decision support system on naive Bayesian classification. *IEEE J. Biomed. Health Inform.* (2015). doi:[10.1109/JBHI.2015.2407157](https://doi.org/10.1109/JBHI.2015.2407157)
32. Lounis, A., Hadjidj, A., Bouabdallah, A., Challal, Y.: Healing on the cloud: secure cloud architecture for medical wireless sensor networks. *Future Gener. Comput. Syst.* (2015). doi:[10.1016/j.future.2015.01.009](https://doi.org/10.1016/j.future.2015.01.009)
33. Mashima, D., Ahamad, M.: Enhancing accountability of electronic health record usage via patient-centric monitoring. In: Proceedings of the 2nd ACM SIGHT International Health Informatics Symposium, IHI '12, pp. 409–418. ACM (2012)
34. Metri, P., Sarote, G.: Privacy issues and challenges in cloud computing. *Int. J. Adv. Eng. Sci. Technol.* **5**, 1–6 (2011)
35. Nabeel, M., Bertino, E.: Privacy preserving delegated access control in public clouds. *IEEE Trans. Knowl. Data Eng.* **26**(9), 2268–2280 (2014)
36. Ning, H., Liu, H., Yang, L.: Cyberentity security in the Internet of things. *Computer* **46**(4), 46–53 (2013)
37. Pecarina, J., Pu, S., Liu, J.C.: SAPPHIRE: anonymity for enhanced control and private collaboration in healthcare clouds. In: Proceedings of the 4th IEEE International Conference Cloud Computing Technology and Science (CloudCom), pp. 99–106. ACM (2012)
38. Puttaswamy, K., Kruegel, C., Zhao, B.: Silverline: toward data confidentiality in storage-intensive cloud applications. In: Proceedings of the 2nd ACM Symposium on Cloud Computing, SOCC '11, pp. 10:1–10:13. ACM (2011)
39. Ruij, S., Stojmenovic, M., Nayak, A.: Privacy preserving access control with authentication for securing data in clouds. In: Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (Ccgrid), CCGRID '12, pp. 556–563. IEEE Computer Society (2012)
40. Shen, Q., Liang, X., Shen, X., Lin, X., Luo, H.Y.: Exploiting geo-distributed clouds for a e-health monitoring system with minimum service delay and privacy preservation. *IEEE J. Biomed. Health Inform.* **18**, 430–439 (2014)

41. Sookhak, M., Gani, A., Talebain, H., Akhunzada, A., Khan, S., Buuya, R., Zomaya, A.: Remote data auditing in cloud computing environments: a survey, taxonomy, and open issues. *ACM Comput. Surv.* **47**(4), 65:1–65:34 (2015)
42. Subashini, S., Kavitha, V.: Review: a survey on security issues in service delivery models of cloud computing. *J. Netw. Comput. Appl.* **34**(1), 1–11 (2011)
43. Sujansky, W., Kunz, D.: A standard-based model for the sharing of patient-generated health information with electronic health records. *Pers. Ubiquit. Comput.* **19**(1), 9–25 (2014)
44. Sundareswaran, S., Squicciarini, A., Lin, D.: Ensuring distributed accountability for data sharing in the cloud. *IEEE Trans. Dependable Secure Comput.* **9**(4), 556–568 (2012)
45. Thilakanathan, D., Chen, S., Nepal, S., Calvo, R., Alem, L.: A platform for secure monitoring and sharing of generic health data in the cloud. *Future Gener. Comput. Syst.* **35**, 102–113 (2014)
46. Thomas, H., Lohr, H., Sadeghi, A., Winandy, M.: Flexible patient-controlled security for electronic health records. In: Proceedings of the 2nd ACM SIGHT International Health Informatics Symposium, IHI '12, pp. 727–732. ACM (2012)
47. Tong, Y., Sun, J., Chow, S., Li, P.: Cloud-assisted mobile-access of health data with privacy and auditability. *IEEE J. Biomed. Health Inform.* **18**(2), 419–429 (2014)
48. U.S. Department of Health & Human Services.: Health Insurance Portability and Accountability Act of 1996 (HIPAA). <http://aspe.hhs.gov/admnsimp/final/pvcpre03.htm> (2015). Accessed 25 April 2015
49. VMware.: Your cloud in healthcare. <http://www.vmware.com/files/pdf/VMware-Your-Cloud-in-Healthcare-Industry-Brief.pdf> (2015). Accessed 22 April 2015
50. Wang, C., Wang, Q., Ren, K., Cao, N., Lou, W.: Toward secure and dependable storage services in cloud computing. *IEEE Trans. Serv. Comput.* **5**(2), 220–232 (2012)
51. Wang, H., Wu, Q., Qin, B., Ferrer, J.: FRR: fair remote retrieval of outsourced private medical records in electronic health networks. *J. Biomed. Inf.* **50**, 226–233 (2014)
52. Wang, Z., Huang, D., Zhu, Y., Li, B., Chung, C.J.: Efficient attribute-based comparable data access control. *IEEE Trans. Comput.* (2015). doi:[10.1109/TC.2015.2401033](https://doi.org/10.1109/TC.2015.2401033)
53. World Health Organization.: E-health. <http://www.who.int/trade/glossary/story021/en/> (2015). Accessed 25 April 2015
54. Wu, R., Ahn, G.J., Hu, H.: Secure sharing of electronic health records in clouds. In: Proceedings of the 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), pp. 711–718 (2012)
55. Xu, L., Cremers, A.: A decentralized pseudonym scheme for cloud-based ehealth systems. In: Proceedings of the 2014 International Conference on Health Informatics, pp. 230–237. ACM (2014)
56. Xu, L., Cremers, A., Wilken, T.: Pseudonymization for secondary use of cloud based electronic health records. Working paper, Academy of Science and Engineering (2015)
57. Yang, J.J., Li, J.Q., Niu, Y.: A hybrid solution for privacy preserving medical data sharing in the cloud environment. *Futur. Gener. Comput. Syst.* **43–44**, 74–86 (2015)
58. Yu, S., Wang, C., Ren, K., Lou, W.: Achieving secure, scalable, and fine-grained data access control in cloud computing. In: Proceedings of the 2010 IEEE INFOCOM Conference, pp. 1–9 (2010)
59. Yu, H.J., Lai, H.S., Chen, K.H., Chou, H.C., Wu, J.M., Dorjgochoo, S., Mendjargal, A., Altangerel, E., Tien, Y.W., Hsueh, C.W., Lai, F.: A sharable cloud-based pancreaticoduodenectomy collaborative database for physicians: emphasis on security and clinical rule supporting. *Comput. Methods Prog. Biomed.* **111**(2), 488–497 (2013)
60. Zhang, R., Liu, L.: Security models and requirements for healthcare application clouds. In: Proceedings of the 3rd IEEE International Conference on Cloud Computing (CLOUD), pp. 268–275 (2010)
61. Zhang, K., Zhou, X., Chen, Y., Wang, X., Ruan, Y.: Sedic: privacy-aware data intensive computing on hybrid clouds. In: Proceedings of the 18th ACM Conference on Computer and Communications Security, CCS '11, pp. 515–526. ACM (2011)

62. Zhang, X., Liu, C., Nepal, S., Pandey, S., Chen, J.: A privacy leakage upper bound constraint-based approach for cost-effective privacy preserving of intermediate data sets in cloud. *IEEE Trans. Parallel Distrib. Syst.* **24**(6), 1192–1202 (2013)
63. Zhang, K., Liang, X., Baura, M., Lu, R., Shen, X.: PHDA: a priority based health data aggregation with privacy preservation for cloud assisted WBANs. *Inf. Sci.* **284**, 130–141 (2014)
64. Zhang, R., Liu, L., Xue, R.: Role-based and time-bound access and management of EHR data. *Secur. Commun. Netw.* **7**(6), 994–1015 (2014)
65. Zhou, J., Cao, Z., Dong, X., Xiong, N., Vasilakos, A.: 4s: a secure and privacy-preserving key management scheme for cloud-assisted wireless body area network in m-healthcare social networks. *Inf. Sci.* **314**, 255–276 (2015)
66. Zhou, J., Lin, X., Dong, X., Cao, Z.: PSMPA: patient self-controllable and multi-level privacy-preserving cooperative authentication in distributed m-healthcare cloud computing system. *IEEE Trans. Parallel Distrib. Syst.* **26**(6), 1693–1703 (2015)
67. Zissis, D., Lekkas, D.: Addressing cloud computing security issues. *Futur. Gener. Comput. Syst.* **28**(3), 583–592 (2012)

Part III

Privacy for Emerging Applications

Chapter 16

Preserving Genome Privacy in Research Studies

Shuang Wang, Xiaoqian Jiang, Dov Fox, and Lucila Ohno-Machado

Abstract As the cost of genome sequencing continues to fall, whole genome sequencing data have become a viable alternative for improving diagnostic accuracy and supporting personalized medicine. Although they have the potential to advance public health and accelerate scientific discoveries, massive collections of genomic data also raise significant concerns about individual privacy. Like traditional clinical information, human genomes may reveal information about individuals (e.g., identity, ethnic group, disease association, predisposition to diseases such as diabetes or cancer, etc.) Even more concerning is the fact that the information is shared with ancestors and descendants, and thus loss of privacy may put the privacy of the entire family at risk. Genome privacy is a big challenge for the entire biomedical community, particularly since scientific discoveries depend on data sharing and obfuscation of data is not a good option to protect privacy. Multiple factors are involved in genomic privacy research. The components that can be used to better protect genome privacy include, but are not limited to, legal, ethical and technical aspects, e.g., federal laws, policies and regulations, informed consent policies, data use agreements, secure data repositories, as well as privacy-preserving data analysis methods. However, genome privacy challenges cannot be addressed by any single component alone. We envision that better privacy protection can be achieved through the incorporation of multiple components. The goal of this chapter to introduce the state-of-the-art in genome privacy research. This chapter begins with an introduction of genome privacy followed by an overview of the legal, ethical and technical aspects of genome privacy. After formalizing the genome privacy

Shuang Wang and Xiaoqian Jiang share the first authorship.

S. Wang (✉) • X. Jiang • L. Ohno-Machado

Department of Biomedical Informatics, University of California, San Diego, CA, USA
e-mail: shw070@ucsd.edu; x1jiang@ucsd.edu; machado@ucsd.edu

D. Fox

School of Law, University of San Diego, San Diego, CA, USA
e-mail: dovfox@sandiego.edu

problem, we will review existing attack models on genomic data. The techniques for mitigating these attacks are discussed. This chapter concludes with the discussion of the challenges and the future directions in genome privacy research.

16.1 Introduction

Biomedical science has been undergoing a significant transformation in the past decade, with an increasing emphasis on data-driven approaches [1]. Rapid advances in genome sequencing technologies have enabled the acquisition of a large amount of genomic data inexpensively. For example, the HiSeq X Ten [2], launched by Illumina in January 2014, delivers the first \$1000 genome at 30 \times coverage. Nowadays, researchers are able to conduct sequencing studies with thousands of individuals [3, 4]. Genome-wide association studies (GWAS) focus on the examination of common genetic variants among different individuals to identify the association between genetic variants and a given trait. As suggested in a recent study [5], a robust detection of novel loci that is likely to be associated with a given disease in GWAS may require even hundreds of thousands of samples. “Big Data” science [6] heavily relies on storage, integration and analysis of the data on an unprecedented scale to facilitate disease prevention (e.g., BRCA mutation test for breast cancer prevention [7]), to improve healthcare (e.g., personalized medicine [8, 9] and precision medicine [10]) and to promote biomedical research (e.g., detecting novel loci for a certain disease [11]).

Because the required sample size in a genomic study can easily overwhelm the availability in a single institution, genomic data sharing plays an important role in promoting biomedical research involving sequencing data. Through data sharing, researchers might also utilize existing genomic data for secondary analysis or even combine genomic data from different sources to increase the statistical power of studies and to accelerate discoveries. In order to make progress toward this goal, NIH launched the database of Genotypes and Phenotypes (dbGaP) [12], which provides open or controlled accesses to genomic data in GWAS depending on the original informed consent from participants. Additionally, NIH also expressed its intention [13] to extend the existing GWAS data sharing policy to cover data from a wider range of genomic studies. However, the increased demand and availability of genomic data also raises serious concerns about the confidentiality of such data and the privacy of individuals [14]. Potential threats to genomic data privacy include, but are not limited to, re-identification risk [15, 16], disease association risk [17], genomic data recovery risk [18], and unintended information disclosure of blood relatives [19]. Over the past few years, many studies have been conducted in the area of genome privacy, which focuses on different aspects of genome privacy, including legal [20], ethics [21, 22] and technology [14, 23].

Genome privacy is emerging as a new interdisciplinary field, which engages researchers from different areas including genomics, bioethics, law, bioinformatics, security, computer science, etc. The recent workshops in genome privacy (i.e., the first and second International Workshop on Genome Privacy [24], the 2014 [25] and the 2015 iDASH Genomic Privacy and Security Challenges [26]) have evidenced the establishment of such a heterogeneous community, in which researchers with different backgrounds can exchange insights, methods, results, legal, social and ethical implications of genome privacy. However, as the co-organizers in the last two workshops, we realize that there exist several technological gaps between the designs of practical genome privacy protection methods and practical genomic applications, primarily due to reduced communications between geneticists and security/privacy experts. On one hand, security or privacy experts overlook the biological context in designing their protection methods. Existing methods [27–29] for protecting the privacy of medical records may not be applicable directly to the genomic data due to the high-dimensional nature of the latter. Perturbation based protection methods [30] applied to genomic data often introduce too much noise and studies based on these noisy outputs are not trustworthy for practical genomic applications. On the other hand, geneticists and bioethicists commonly lack the necessary knowledge in formalizing the attack models or articulating a tolerable level of privacy risk. Therefore, further interaction and cooperation of these disciplines is necessary. The goal of this chapter is to summarize and systematize knowledge of genome privacy.

The rest of the chapter is organized as follows. Section 16.2 discusses genome privacy in terms of policies and legal regulations in the United States. Next, Sect. 16.3 focuses on information technology that is related to genomic privacy, including risk models and protection mechanisms that have been proposed. Last, Sect. 16.4 provides conclusions, discussion, and future directions.

16.2 Policies, Legal Regulation and Ethical Principles of Genome Privacy

In this section, we will present current policies, legal regulations and ethical principles regarding genomic data sharing and privacy protection in the United States.

16.2.1 NIH Policies for Genomic Data Sharing

16.2.1.1 GWAS Data Sharing Policy

In August 2007, the National Institutes of Health (NIH) announced the policy for sharing of data obtained in NIH-supported Genome-Wide Association Studies

(GWAS policy) [31]. This policy introduced two tiers of GWAS data sharing models, i.e., *unrestricted* and *controlled* access. Furthermore, NIH created a central repository [i.e., the database of Genotypes and Phenotypes (dbGaP)] [12] to archive and distribute GWAS data. The above initiatives aimed at promoting scientific discovery and maximizing the public benefits by sharing GWAS data in compliance with the original informed consents from participants, taking into consideration privacy risks. For example, a recent report [32] analyzed 304 dbGaP studies deposited before December 1, 2013, where 17,746 data access requests (with an approval rate of 68 %) had been submitted by 2221 researchers and 6800 collaborators from 41 countries. The secondary use of dbGaP data resulted in a total of 924 publications and significant breakthroughs.

Data Submission to dbGaP To protect the privacy of participants, individual-level data are de-identified by the investigators before being submitted to dbGaP. The submitted data are organized based on the participants' informed consent, where data in the same consent group have the same data use limitations (e.g., limitations to a specific disease or the embargo date for a secondary research). Then, each study in dbGaP is assigned with a unique accession number and associated with additional information, which includes, but is not limited to dataset description, data use certification, data use limitation, related publication, funding source, etc.

Data Access to dbGaP Researchers can download and use unrestricted datasets (i.e., public datasets) without authorization and time limits. In contrast, any request to the controlled-access (i.e., restricted) dataset from researchers must be first reviewed by the NIH data access committee (DAC), which is composed of experts in bioethics and relevant scientific areas. Investigators need to agree to the terms elaborated in the data use certification for each requested dataset, and to meet data security requirements in order to manage dbGaP data. A DAC will approve or disapprove a given request by evaluating the consistency between the proposed secondary research and data use limitations of the requested dataset. Currently, NIH has established 16 DACs representing their 18 NIH Institutes and Centers (ICs), and the average processing time of a request is 14 days. Recently, NIH released a statement related to the use of cloud computing services for storage and analysis of controlled-access data [33]. Although cloud computing provides better scalability and less management effort than local clusters, ensuring data security and privacy in cloud environments is still the responsibility of the researchers and their institutions.

16.2.1.2 Genomic Data Sharing Policy

Following on the success of the GWAS policy and due to the increasing volume and complexity of genomic data (e.g., non-human genomic data, gene expression data, whole genome sequencing data, etc.), NIH developed a new Genomic Data Sharing (GDS) policy [13] in 2014 to meet the needs for managing and distributing large scale genomic data from a wide range of genomic research. Similar to the GWAS policy, the GDS policy allows genomic data to be shared through either

Table 16.1 Key aspects of GWAS and genomic data sharing (GDS) policies

	GWAS policy	GDS policy
Effective dates	January 25, 2008	January 25, 2015
Scope	Human GWAS data	Both human and non-human genomic data
Consent standard for data collected before the effective date	<ul style="list-style-type: none"> • If consent is not available, data may still be submitted to NIH • If consent is available, IRB documents are required to ensure consistency 	Same
Consent standard for data collected after the effective date	Specific discussion about sharing participants' genotype and phenotype data for research purposes within the informed consent is expected by NIH	<ul style="list-style-type: none"> • Genomic data, clinical specimens and cell lines should be consented for research use with explicit statement of either open or controlled access • Expectations might be allowed upon special request
Data submission timeline	As soon as the data quality control has been finished	<ul style="list-style-type: none"> • Not expected for raw data or initial sequence reads • Within 3 months for processed data (e.g., aligned sequences, SNPs, etc.) • As soon as analyses are completed for final analysis data (e.g., genotype-phenotype data)
Data release timeline	Immediate data release, but with a 12 month publication embargo for secondary use	<ul style="list-style-type: none"> • N/A for raw data or initial sequence reads • 6 months for non-human data, and processed data • Immediate release of final analysis data

open or controlled access depending on the nature of informed consent [34]. The controlled access data may be available for secondary use only after researchers have obtained approval from the NIH for the particular study. However, there are also several important changes in the GDS policy. The key differences between GWAS and GDS policies are summarized in Table 16.1.

16.2.2 U.S. Legal Regulations for Genomic Data

The U.S. Supreme Court has not established a constitutional right to informational privacy that it has said applies to whole genome sequence data. Most recently in *Maryland v. King*, a majority of the Supreme Court authorized police under the Fourth Amendment bar on unreasonable searches and seizures to take DNA from anyone they arrest, so long as the only part of the DNA they analyze does not encode instructions for making proteins [35]. The constitutional logic of that decision is that DNA may be extracted by law enforcement if it does not reveal the kinds of sensitive genetic or medical information that the Court referred to as “more far-reaching and complex characteristics like genetic traits” [36].

The most significant source of medical privacy protection beyond the policing context is a congressional law called the Health Information Portability and Accountability Act, or HIPAA [37]. Passed in 1996, HIPAA establishes rules designed to secure the privacy of personally identifiable health information. Those rules are then interpreted by the Department of Health and Human Services (HHS). HIPAA’s Privacy Rule governs the conditions under which certain “covered entities” such as healthcare providers can disclose patient-identifiable information. In 2013, the Health Information Technology for Economic and Clinical Health (HITECH) Act added the business associates of any previously covered entities to those liable for the disclosing protected health information under HIPAA [38].

Specifically, these entities are forbidden from disclosing patient-identifiable information (unless with patient consent, or to HHS in the event of a compliance investigation or enforcement action). This includes the patient’s name, address, social security number, and “biometric identifier” or “any other unique identifying number, characteristic, or code” [39]. HHS has not clarified whether genetic or genomic information counts among these protected patient-identifiable forms of health information. Importantly, HIPAA imposes no restrictions on the use or disclosure of health information that does not identify an individual or provide a reasonable basis with which to identify him or her [40].

Institutions that do not count as “covered entities” subject to the requirements of HIPAA and HITECH must instead follow the so-called Common Rule [41]. This requires human subjects’ research funded by the federal government (state-funded or open-source genomics projects are not covered) to obtain informed consent, minimize risks to participants, and acquire satisfactory marks by an institutional review board (IRB). Free from these Common Rule requirements, however, is research that uses data where the identity of the subject cannot be readily determined [42].

The “de-identified” data can be used for research purposes without meeting the Common Rule requirements. This implies that participants in federally funded genomics studies will be not given additional consent or be advised about the newly discovered risk of re-identification. This has recently led HHS to propose new rules that better address the risks of de-identified data that can be re-identified, as presented in Fig. 16.1.

Rapidly evolving advances in technology coupled with the increasing volume of data readily available may soon allow identification of an individual from data that is currently considered de-identified . . . We are considering adopting the HIPAA standards for purposes of the Common Rule regarding what constitutes individually identifiable information . . . it might be advisable to evaluate the set of identifiers that must be removed for a data set to be considered “de-identified” under both human subjects regulations and the HIPAA Privacy Rule . . . we are considering categorizing all research involving the primary collection of biospecimens as well as storage and secondary analysis of existing biospecimens as research involving identifiable information [43].

Fig. 16.1 HHS’s new rules to address the risks of de-identified data that can be re-identified

At present, however, “de-identified” data remains unprotected under HIPAA if all 18 HIPAA identifiers are removed or an expert determines that the risk of re-identification is negligible [44]. Courts have determined that either identifier removal or expert determination, but not both, is required to qualify data as de-identified, and thus unprotected under the HIPAA Privacy Rule [45].

The Genetic Information Nondiscrimination Act, known as GINA, builds on HIPAA protections, though it does not directly regulate privacy concerns related to access to genomic data [20]. Instead this 2008 law makes it illegal for employers with more than 15 employees to fire or refuse to hire workers, or “otherwise to discriminate against any employee with respect to the compensation, terms, conditions, or privileges of employment”, based on their genetic test results or family history of disease [46]. GINA also creates a cause of legal action for people to sue health insurance companies for investigating their genetic information or using it to set their premiums or deny them coverage [47]. Most GINA plaintiffs failed, however, for difficulties identifying the misuse of genetic as opposed to other kinds information [48]. While the law applies only to asymptomatic individuals, and not those who have “manifested disease”, the 2010 Patient Protection and Affordable Care Act (ACA) prohibits health insurers from setting premiums or determining eligibility for coverage based on signs and symptoms of genetic disease.

GINA extends coverage to the genetic information of family members related to individuals, including disease manifestation, but only if relevant to determining the individual’s risk for a genetic disease [49]. Courts have treated as “genetic”, and thus eligible for GINA protection, “information about whether an employee’s ‘primary relative’ has a history of prostate cancer” [50], but not a plaintiff’s wife’s diagnosis of multiple sclerosis [51]. GINA’s protections do not cover the disadvantaging use of genetic information in disability insurance, life insurance, or long-term care. Nor does the law apply to athletic programs or health benefits for federal employees, members of the military, veterans seeking healthcare through the Department of Veterans Affairs, or the Indian Health Service [52].

GINA sets threshold federal protections in a way that permits states to provide additional safeguards [53]. Nineteen states add some form of such additional protections; these vary widely from state to state. Ten states have general privacy

protection laws that require informed consent to collect, access, retain or disclose the genetic information and that establish rules on collection, access, retention and disclosure of such information [53]. Among the broadest such protections are offered in California, which prohibits genetic discrimination in housing, receipt of services, qualifications for licensing, emergency medical services, and participation in any state-funded programs [54]. The five states like Alaska that afford certain property rights tend to limit those rights to the DNA sample itself, with limited protection for the genetic information it confers [55]. Some of these statutes also set forth added informed consent requirements to collect or disclose DNA samples and give people a private right of action against those who collect a DNA sample or disclose DNA testing results without such content [56].

16.2.3 Ethical Principles for Genome Privacy

There are a number of risks to preserving the anonymity of genetic information from individuals. Genomic data can be leaked by a reckless clinician or sequencing facility, or de-anonymized by a hacker or disgruntled employee. Because the genomes of closely related people are so similar, the disclosure of one person's genome—whether voluntary, accidental, or malicious—can reveal sensitive information about his or her siblings or children too [19].

Serious concerns about individual privacy follow from the risk that genomic data might be connected to the person they come from. This can be used to establish parental status in family disputes, to identify suspected criminals in police searches, or to disclose health risks of possible interest to insurers or employers. Dr. Nornalane Lindor of the Mayo Clinic tells the story of a cancer patient, for example, whose grandchild was rejected in her application to become a U.S. army helicopter pilot when it was revealed that she had been genetically tested for a cancer mutation [57].

Informed consent and data use agreements (DUAs) play important roles in controlled access for data sharing. Informed consent from a subject describes whether or to what extent the data are appropriate for future research use. In addition, an individual participant is also able to request removal of his or her data from future research use by withdrawing the consent [58]. A DUA regulates the responsibilities of investigators including, but not limited to, the type of research, expiration of the contract, liabilities of protecting data confidentiality, reporting any violation, and not attempting to re-identify participants.

Informed consent to participate in a genome sequencing study should at a minimum make clear how the data will be used now and in the future, the degree to which participants can control future use of the data, and potential known and unknown benefits and risks of participation, including whether they will be informed of incidental findings. People may assess in very different ways the desire to learn about genetic conditions for especially late-onset disorders or disease [59].

To increase participation, some genomics data studies have used opt-out models and blanket consent policies that do not limit the scope and duration of consent,

but these threaten to under-inform participants about the relevant risks and benefits. As mentioned previously, the Common Rule formally regulates federally funded human subjects research. But it does not require IRB review or additional consent for studies using data for which HIPAA identifiers are removed. So if the initial consent provided to participate in a genomic data study did not include information about privacy risks or when the data can be used for different studies, a new consent need not be obtained unless state laws require it.

In light of re-identification risks, a number of major studies, including the HapMap and 1000 Genomes project, undertook procedures to acquire updated consent from participants [60]. The “re-consent” process included informing those participants who were still alive about the inability to guarantee privacy and removal of their information without their newly apprised agreement. The Personal Genome Project (PGP) has adopted an “open consent” policy that makes no privacy guarantees. Instead, it uses privacy education and ongoing participation in order to cultivate participants willing to share their genomic information with limited control to review and add data [61]. The PGP’s open consent approach updates consent for genomics data studies in ways that facilitate participant education and transparency about re-identification risks [62].

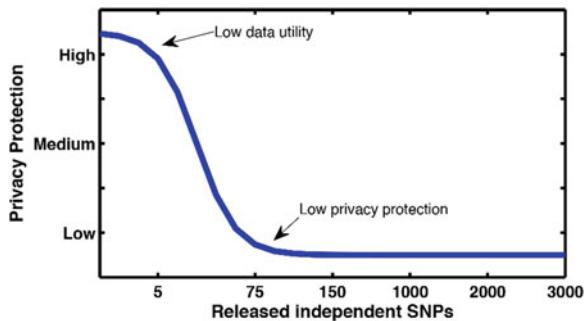
16.2.4 Summary

A patchwork of legal and regulatory mechanisms helps to protect against genetic discrimination and to address privacy concerns. GINA prohibits genetic discrimination in the health insurance and employment contexts. But it does not apply to disability insurance, life insurance, or long-term care. Nor does it regulate use, access, security, or disclosure of genetic data. HIPAA addresses such privacy concerns, but that law does not make clear whether it covers genomic information. Nor do its protections extend to certain academic institutions, federal agencies or scientific consortia that tend to manage public genomic projects. While these entities are bound by the requirements of the Common Rule, those privacy protections are less strict than HIPAA, at least in advance of the proposed updates mentioned earlier. State laws provide additional protections in some cases, but no systemic coverage.

16.3 Information Technology for Genome Privacy

In this section, we focus on information technology to protect genome privacy. This section starts with the discussion about existing genome privacy risks followed by the review of genome privacy protection technologies. Community efforts on genome privacy protection are discussed at the end of this section.

Fig. 16.2 Privacy protection and number of released independent single nucleotide polymorphisms (SNPs) base on the report in [63]



16.3.1 Genome Privacy Risks

As shown in Fig. 16.2, Lin et al. [63] showed that as few as 30–80 statistically independent Single Nucleotide Polymorphisms (SNPs) are sufficient to uniquely identify a single person. Genome privacy issues become even more critical because recent research indicates that even aggregated genome data (e.g., test statistics), which were assumed to be low risk, can be used for re-identifying an individual or reconstructing original SNP sequences [18].

Figure 16.3 illustrates the famous Homer's attack [15], where the existence of a person of interest within a complex DNA mixture (e.g., a case group with a certain disease) can be asserted with high confidence through comparisons of allele frequencies of the reference population, the individual, and the mixture, followed by a statistical test.

Because of the potential privacy risks reported in previous studies, the NIH has removed most aggregated results from the public domain [31]. As whole genome sequencing is becoming more affordable for many individuals, thorny privacy and ethical issues must be carefully addressed [64]. A recent paper by Naveed et al. reviewed various privacy attacks on genomic data and contextualized these attacks from the perspective of medicine and public policy [62]. Another review paper by Erlich and Narayanan also summarized routes for breaching and protecting genetic privacy [65]. In this chapter, we will focus on the most recent technological progress on protection of human genome privacy in different application scenarios.

16.3.2 Genome Privacy Protection Technologies

To contextualize our findings, in what follows we group the technological solutions into two categories based on their focus: (1) *protection of the computation process of genomic data analysis*, and (2) *protection of raw data or of the outcome of genomic data analysis*.

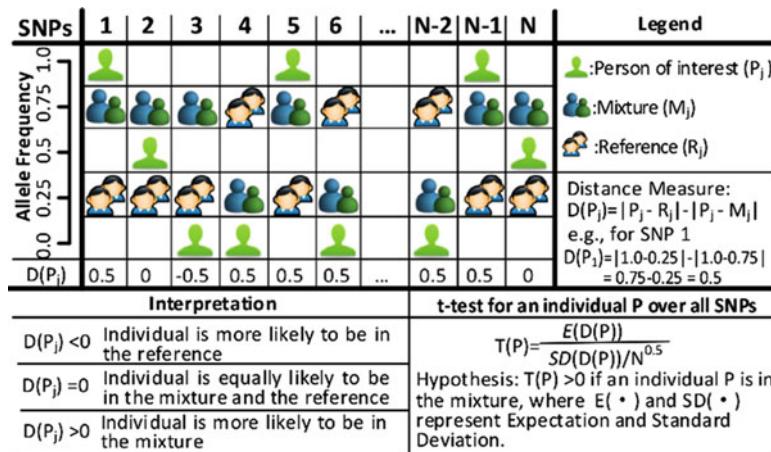


Fig. 16.3 Illustration of Homer's attacks, where $|P_j - R_j|$ and $|P_j - M_j|$ measure how the person's allele frequency P_j differs from the allele frequencies of the reference population and the mixture, respectively. By sampling a large number of N SNPs, the distance measure $D(P_j)$ will follow a normal distribution due to the central limit theorem. Then, a one-sample t-test for this individual over all sampled SNPs can be used to verify the hypothesis that an individual is in the mixture ($T(P) > 0$)

The first category of methods aims at ensuring the privacy and security during the computation process of genomic data analyses. Lauter et al. developed homomorphic encryption methods to support private computation on encrypted genomic data [66–68], which allow computation to be carried out on ciphertext (encrypted data), thus generating an encrypted result that, when decrypted, matches the result of operations performed on the plaintext (original, non-encrypted data) [69].

Homomorphic encryption methods are highly generalizable and are promising for secure outsourcing of computation [70], such as computing in public clouds. However, homomorphic encryption methods are computation and storage intensive. For task-specific scenarios, more efficient solutions have been proposed. Ayday et al. developed a privacy-preserving mechanism for processing raw genomic data to enable a medical unit to privately retrieve a subset of the short reads of the patients (which include a definite range of nucleotides depending on the type of the genetic test) without revealing the nature of the genetic test to the encrypted biobank [71].

Huang et al. developed a new tool, GenoGuard, based on honey encryption (i.e., a process that makes it hard for attackers to check whether they obtained the desired data) to provide strong protection for genomic data against brute-force attacks [72]. Danezis and Cristofaro proposed protocols for privacy-preserving disease susceptibility testing [73]. Djamico et al. developed a secure evaluation protocol to calculate Warfarin dosage [74]. Lu et al. proposed an efficient protocol for secure outsourcing of Genome-Wide Association studies (GWAS) [75]. Verle et al. developed a privacy-preserving GWAS statistical analysis method using exact

logistic regression [76]. These protocols are customized for certain data analysis tasks and allow multiple parties to collaborate via secure multiparty computing (SMC).

Kantarcioğlu et al. developed a cryptographic approach to share and query genomic sequences [77]. In these cases, participating parties jointly compute a function over their inputs, and keep these inputs private. Each party can perform certain computations locally over the controlled-access (private) data, and will only exchange intermediary results to synthesize a global model. These protocols often require synchronization and involve a large amount of peer-to-peer communication.

The second category of methods aims at protecting raw data or the outcome of genomic data analysis to mitigate privacy risks. Malin developed a genomic sequence anonymization technique with generalized lattices based on the privacy protection schema of k -anonymity [28], which makes it hard to learn features that distinguish one genetic sequence from $k - 1$ other entries in a collection [78].

Loukides et al. proposed an anonymization model to protect potentially linkable clinical features and modified them in a way that they could no longer be used to link a genomic sequence to a small number of patients, while preserving the associations between genomic sequences and specific sets of clinical features corresponding to GWAS-related diseases [79].

Many recent studies adopt differential privacy [27], a strong and provable privacy criteria, to formally quantify what an adversary could learn from released information. Yu et al. developed differentially private logistic regression for detecting multiple-SNP association in GWAS databases [80]. Wang et al. developed a differentially private genomic data dissemination model through top-down specialization [81]. Johnathon and Shmatikov proposed a privacy-preserving mechanism to conduct chi-squared tests on genome data [82]. Uhler et al. developed an alternative solution for differentially private chi-squared test, as well as a way to release the M most relevant SNPs with respect to a specific phenotype [83]. Yu et al. improved the previous methodology in releasing top M most relevant SNPs with more favorable privacy and utility tradeoffs, and provided formal proofs [84, 85].

16.3.3 Community Efforts on Genome Privacy Protection

The challenges in genomic privacy gave rise to a new research community [86]. Privacy and security researchers have successfully hosted the International Workshop on Genome Privacy and Security (GenoPri) 2 years in a row to cover a wide range of topics from access control to privacy enhancing technologies for genomic data. The latest workshop (GenoPri'15) [87] was held in conjunction with the 36th IEEE Symposium on Security and Privacy (S&P) and brought together a highly interdisciplinary community involving all aspects of genome privacy and security research to jointly tackle the emerging challenges.

From the biomedical side, the National Center for Biomedical Computing hosted at the University of California San Diego, iDASH (integrating Data for Analysis,

‘anonymization’ and SHaring) [88] is dedicated to bridge the gap between the privacy protection methodology and real genomic data analysis needs. In 2014, iDASH hosted a competition for Critical Assessment of Data Privacy and Protection (CADPP) as a community effort to assess existing differential privacy methods and facilitate the development of new models for privacy-preserving genomic data sharing and analysis [25]. The competition revealed that differential privacy-based data perturbation techniques have several limitations for sharing a large volume of human genomic data, but that they could be used to disseminate analytical results (using GWAS-like statistics), through services like those offered by NCBI [89].

In 2015, iDASH engaged a larger body of the privacy and security community to host the second CADPP competition to assess the capacity of state-of-the-art cryptographic technologies like Homomorphic encryption (HME) and secure multi-party computing (SMC) in the context of secure and privacy-preserving data outsourcing and distributed computing using genomic data [70]. We designed one challenge for secure Hamming and Edit distance computation. The Hamming distance measures the number of pair-wise differences between two equal-length sequences. The Edit distance is defined as the minimum number of operations (e.g., insertion, deletion, substitution, etc.) required to transform one genome sequence into the other. Both distances are widely used to measure the similarity between genome sequences.

The results were very promising, as Hamming and Edit distance approximations over two sequences of 100k length could be calculated within 10 min on encrypted data and secure collaboration over the Internet for Hamming and Edit distance approximations of 100K could be conducted in 20 min. On the other hand, the competition also showed that a full-fledged GWAS could still not be efficiently done on encrypted DNA sequences, and that multi-gigabytes of memory or Internet transmission are required to enable HME or SMC, making it hard to scale up analyses based on these techniques.

16.4 Conclusion

In this chapter, we discussed three facets (policy, ethics, and information technology) of genomic data privacy. The policy, ethical, and technical issues surrounding genomic privacy research can be summarized as a balance between maximizing benefit and minimizing harm. Policy-level protection provides a system of rules to govern behavior in conducting genome research, by trying to establish an atmosphere of basic trust and an anticipation of security. However, policies and regulations alone are not sufficient to protect genome privacy. Researchers who have access to sensitive genomic data should act in an ethical manner by following informed consent documents and data use agreements. Complying with current legal and privacy regulations, as well as following ethical principles, are essential for further progress of genomic research. However, information technology is also indispensable in securing data during storage, sharing and analyses based

on state-of-the-art data encryption and anonymization methods. We conclude that current challenges and limitations in genome privacy cannot be fully solved by any single facet. We envisage that genome privacy protection can be improved through continuous collaborations among policy makers, ethics advocates, and researchers.

Acknowledgements This work was funded by NHGRI (K99HG008175), NLM (R00LM011392, R21LM012060), and NHLBI (U54HL108460).

References

1. Howe, D., Costanzo, M., Fey, P., et al.: Big data: the future of biocuration. *Nature* **455**, 47–50 (2008). <http://dx.doi.org/10.1038/455047a>. Accessed 11 Jul 2014
2. HiSeq X Ten.: 1000 dollar genome sequencing. <http://www.illumina.com/systems/hiseq-x-sequencing-system.ilmn>. Accessed 11 Jul 2014
3. Abecasis, G.R., Auton, A., Brooks, L.D., et al.: An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012). doi:[10.1038/nature11632](https://doi.org/10.1038/nature11632)
4. Fu, W., O'Connor, T.D., Jun, G., et al.: Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013). doi:[10.1038/nature11690](https://doi.org/10.1038/nature11690)
5. Park, J.-H., Wacholder, S., Gail, M.H., et al.: Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575. (2010). doi:[10.1038/ng.610](https://doi.org/10.1038/ng.610)
6. Marx, V.: Biology: the big challenges of big data. *Nature* **498**, 255–260 (2013). doi:[10.1038/498255a](https://doi.org/10.1038/498255a)
7. Bradbury, A.R., Dignam, J.J., Ibe, C.N., et al. How often do BRCA mutation carriers tell their young children of the family's risk for cancer? a study of parental disclosure of BRCA mutations to minors and young adults. *J. Clin. Oncol.* **25**, 3705–3711 (2007). doi:[10.1200/JCO.2006.09.1900](https://doi.org/10.1200/JCO.2006.09.1900)
8. Willard, H.F., Angrist, M., Ginsburg, G.S.: Genomic medicine: genetic variation and its impact on the future of health care. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **360**, 1543–1550 (2005). doi:[10.1098/rstb.2005.1683](https://doi.org/10.1098/rstb.2005.1683)
9. Pulley, J.M., Denny, J.C., Peterson, J.F., et al.: Operational implementation of prospective genotyping for personalized medicine: the design of the Vanderbilt PREDICT project. *Clin. Pharmacol. Ther.* **92**, 87–95 (2012). doi:[10.1038/clpt.2011.371](https://doi.org/10.1038/clpt.2011.371)
10. Collins, F.S., Varmus, H.: A new initiative on precision medicine. *N. Engl. J. Med.* **372**, 793–795 (2015)
11. Visscher, P.M., Brown, M.A., McCarthy, M.I., et al.: Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012). doi:[10.1016/j.ajhg.2011.11.029](https://doi.org/10.1016/j.ajhg.2011.11.029)
12. Mailman, M.D., Feolo, M., Jin, Y., et al.: The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181–1186 (2007). doi:[10.1038/ng1007-1181](https://doi.org/10.1038/ng1007-1181)
13. NIH Genomic Data Sharing Policy.: <http://gds.nih.gov/03policy2.html> (2014)
14. Lin, Z., Owen, A.B., Altman, R.B.: Genetics. Genomic research and human subject privacy. *Science* **305**, 183 (2004). doi:[10.1126/science.1095019](https://doi.org/10.1126/science.1095019)
15. Homer, N., Szelinger, S., Redman, M., et al.: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4**, e1000167 (2008)
16. Gymrek, M., McGuire, A.L., Golan, D., et al.: Identifying personal genomes by surname inference. *Science* **339**, 321–324 (2013)
17. Nyholt, D.R., Yu, C.-E., Visscher, P.M.: On Jim Watson's APOE status: genetic information is hard to hide. *Eur. J. Hum. Genet.* **17**, 147–149 (2009). doi:[10.1038/ejhg.2008.198](https://doi.org/10.1038/ejhg.2008.198)

18. Wang, R., Li, Y.F., Wang, X., et al.: Learning your identity and disease from research papers. In: Proceedings of the 16th ACM Conference on Computer and Communications Security - CCS '09, vol. 534. ACM Press, New York (2009). doi:[10.1145/1653662.1653726](https://doi.org/10.1145/1653662.1653726)
19. Humbert, M., Ayday, E., Hubaux, J.-P., et al.: Addressing the concerns of the Lacks family: quantification of kin genomic privacy. In: Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security - CCS '13, pp. 1141–1152. ACM Press, New York (2013). doi:[10.1145/2508859.2516707](https://doi.org/10.1145/2508859.2516707)
20. Genetic Information Nondiscrimination Act.: (2008), <http://www.eeoc.gov/laws/statutes/gina.cfm>. Accessed 11 Jul 2014
21. McGuire, A.L., Caulfield, T., Cho, M.K.: Research ethics and the challenge of whole-genome sequencing. *Nat. Rev. Genet.* **9**, 152–156 (2008). doi:[10.1038/nrg2302](https://doi.org/10.1038/nrg2302)
22. Caulfield, T., McGuire, A.L., Cho, M., et al.: Research ethics recommendations for whole-genome research: consensus statement. *PLoS Biol.* **6**, e73 (2008). doi:[10.1371/journal.pbio.0060073](https://doi.org/10.1371/journal.pbio.0060073)
23. Sankararaman, S., Obozinski, G., Jordan, M.I., et al.: Genomic privacy and limits of individual detection in a pool. *Nat. Genet.* **41**, 965–967 (2009). <http://dx.doi.org/10.1038/ng.436>. Accessed 18 Apr 2014
24. Amsterdam Workshop on Genome Privacy. <http://seclab.soic.indiana.edu/GenomePrivacy> (2014)
25. 2014 iDASH Genome Privacy Protection Challenge Workshop. <http://www.humangenomeprivacy.org/2014> (2014)
26. 2015 iDASH Privacy and Security Workshop. <http://www.humangenomeprivacy.org/2015/>. Accessed 02 Jan, 2015
27. Dwork, C.: Differential privacy. *Int. Colloq. Autom. Lang. Program.* **405**, 2:1–2:12 (2006)
28. Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Uncertainty Fuzziness Knowl.-Based Syst.* **10**, 557–570 (2002)
29. Li, N., Li, T., Venkatasubramanian, S.: t closeness?: privacy beyond k-anonymity and - diversity. In: IEEE 23rd International Conference on Data Engineering, pp. 106–115. IEEE (2007). <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4221659>
30. Yu, F., Fienberg, S.E., Slavkovic, A.B., et al.: Scalable privacy-preserving data sharing methodology for genome-wide association studies. *J. Biomed. Inform.* Published Online First: 6 February (2014). doi:[10.1016/j.jbi.2014.01.008](https://doi.org/10.1016/j.jbi.2014.01.008)
31. Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS) (2007). <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html>. Accessed 11 Jul 2014
32. Committees the NI of HGDSC.: Data use under the NIH GWAS data sharing policy and future directions. *Nat. Genet.* **46**, 934–938 (2014). <http://dx.doi.org/10.1038/ng.3062>
33. NIH security best practices for controlled-access data subject to the NIH genomic data sharing (GDS) policy. http://www.ncbi.nlm.nih.gov/projects/gap/pdf/dbgap_2b_security_procedures.pdf. Accessed 20 Mar 2015
34. Dondorp, W.J., de Wert, G.M.W.R.: The ‘thousand-dollar genome’: an ethical exploration. *Eur. J. Hum. Genet.* **21**:S6–S26 (2013)
35. Maryland v. King. S. Ct. 2013;133:1958
36. Maryland v. King. S. Ct. 2013;133:1967
37. Health Insurance Portability and Accountability Act (HIPAA). <http://www.hhs.gov/ocr/hipaa>. Accessed 11 Jul 2014
38. New rule protects patient privacy, secures health information. U.S. Department of Health and Human Services. <http://www.hhs.gov/news/press/2013pres/01/20130117b.html>. Accessed 11 Jul 2014
39. HIPAA Privacy Rule, 45 C.F.R. § 164 (2014)
40. Nass, S.J., Levit, L.A., Gostin, L.O.: Beyond the HIPAA privacy rule: enhancing privacy, improving health through research. The National Academies Press, Washington, DC (2009)
41. Federal policy for the protection of human subjects. U.S. Department of Health and Human Services. <http://www.hhs.gov/ohrp/humansubjects/commonrule/>. Accessed 12 Mar 2015

42. 45 C.F.R. § 46.101(b)(4)
43. Human Subject Research Protections, 76 Fed. Reg. 44,512, 44,524–25 (July 26, 2011)
44. 45 C.F.R. § 160.103, 164.514, 164.514
45. Baser v. Dep’t of Veterans Affairs, 2014 U.S. Dist. LEXIS 137602, at *11 (E.D. Mich. Sept. 30, 2014); Steinberg v. CVS Caremark Corp., 899 F. Supp. 2d 331, 336 (E.D. Pa. 2012)
46. 42 U.S.C. § 2000ff
47. 29 U.S.C. § 1182
48. E.g., Dumas v. Hurley Med. Ctr., 837 F. Supp. 2d 655, 659 (E.D. Mich. 2011); Bell v. PSS World Med., Inc., 2012 U.S. Dist. LEXIS 183288 (M.D. Fla. Dec. 7, 2012); Culbreth v. Wash. Metro. Area Transit Auth., 2012 U.S. Dist. LEXIS 37335 (D. Md. Mar. 19, 2011)
49. 42 U.S.C. § 2000ff(3)
50. Lee v. City of Moraine Fire Dep’t, 2014 U.S. Dist. LEXIS 61385, at *16 (S.D. Ohio May 2, 2014)
51. Poore v. Peterbilt of Bristol, L.L.C., 852 F. Supp. 2d 727, 730–31 (W.D. Va. 2012)
52. Slaughter, L.: Genetic information non-discrimination act. *Harv. J. Legis.* **50**, 41 (2013)
53. For the study of bioethical issues PC. Privacy and progress in whole genome sequencing (2012)
54. California Genetic Information Nondiscrimination Act (2011). <http://geneticprivacynetwork.org/about-calgina/>. Accessed 11 Jul 2014
55. Alaska Genetic Information Nondiscrimination Act (2014). http://doa.alaska.gov/dop/fileadmin/Equal_Employment/pdf/EEOP_Policy_Statement.pdf. Accessed 11 Mar 2015
56. Prince, A.E.R.: Comprehensive protection of genetic information. *Brooklyn Law Rev.* **79**, 175–227 (2013)
57. Lindor, N.M.: Personal autonomy in the genomic era. In: Video Proceedings of Mayo Clinic Individualizing Medicine Conference (2012)
58. Khan, A., Capps, B.J., Sum, M.Y., et al.: Informed consent for human genetic and genomic studies: a systematic review. *Clin. Genet.* **86**, 199–206 (2014)
59. Wolf, S.M., Crock, B.N., Van Ness, B., et al.: Managing incidental findings and research results in genomic research involving biobanks and archived data sets. *Genet. Med.* **14**, 361–384 (2012)
60. Rodriguez, L.L., Brooks, L.D., Greenberg, J.H., et al.: The complexities of genomic identifiability. *Science* **339**, 275–276 (2013)
61. Ball, M.P., Bobe, J.R., Chou, M.F., et al.: Harvard personal genome project: lessons from participatory public research. *Genome Med.* **6**, 10 (2014)
62. Naveed, M., Ayday, E., Clayton, E.W., et al.: Privacy and security in the genomic era. Published Online First: 8 May 2014. <http://arxiv.org/abs/1405.1891>. Accessed 11 Aug 2014
63. Lin, Z., Owen, A.B., Altman, R.B. Genetics. Genomic research and human subject privacy. *Science* **305**, 183 (2004). doi:[10.1126/science.1095019](https://doi.org/10.1126/science.1095019)
64. Ayday, E., De Cristofaro, E., Hubaux, J.-P., et al. Whole genome sequencing: revolutionary medicine or privacy nightmare? Computer (Long Beach Calif) **48**, 58–66 (2015). doi:[10.1109/MC.2015.59](https://doi.org/10.1109/MC.2015.59)
65. Erlich, Y., Narayanan, A.: Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.* **15**, 409–421 (2014). doi:[10.1038/nrg3723](https://doi.org/10.1038/nrg3723)
66. Lauter, K., Lopez-Alt, A., Naehrig, M.: Private computation on encrypted genomic data. In: 14th Privacy Enhancing Technologies Symposium, Workshop on Genome Privacy. <http://seclab.soic.indiana.edu/GenomePrivacy/papers/Genome%20Privacy-paper9.pdf>. (2014). 29 July 2014, date last accessed
67. Bos, J.W., Lauter, K., Naehrig, M.: Private predictive analysis on encrypted medical data. *J. Biomed. Inform.* **50**, 234–243 (2014). doi:[10.1016/j.jbi.2014.04.003](https://doi.org/10.1016/j.jbi.2014.04.003)
68. Cheon, J.H., Kim, M., Lauter, K.: Homomorphic computation of edit distance. In: WAHC’15 - 3rd Workshop on Encrypted Computing and Applied Homomorphic Cryptography (2015)
69. Homomorphic_Encryption.: http://en.wikipedia.org/w/index.php?title=Homomorphic_encryption&oldid=653811034 (2015). Accessed 29 Mar 2015
70. Check Hayden, E.: Cloud cover protects gene data. *Nature* **519**, 400–401 (2015). doi:[10.1038/519400a](https://doi.org/10.1038/519400a)

71. Ayday, E., Raisaro, J.L., Hengartner, U., et al.: Privacy-preserving processing of raw genomic data. *Data Priv. Manag. Auton. Spontaneous Secur.* **8247**, 133–147 (2014). <http://infoscience.epfl.ch/record/187573>. Accessed 31 Mar 2015
72. Huang, Z., Ayday, E., Fellay, J., et al.: GenoGuard: protecting genomic data against brute-force attacks. In: 36th IEEE Symposium on Security and Privacy (S&P 2015), San Jose (2015). <http://infoscience.epfl.ch/record/206772>. Accessed 31 Mar 2015
73. Danezis, G.: Simpler protocols for privacy-preserving disease susceptibility testing. In: 14th Privacy Enhancing Technologies Symposium, Workshop on Genome Privacy (GenoPri'14), Amsterdam (2014)
74. Djatmiko, M., Friedman, A., Boreli, R., et al.: Secure evaluation protocol for personalized medicine. In: 14th Privacy Enhancing Technologies Symposium, Workshop on Genome Privacy (GenoPri'14), Amsterdam (2014)
75. Lu, W., Yamada, Y., Sakuma, J.: Efficient secure outsourcing of genome-wide association studies. In: 2nd International Workshop on Genome Privacy and Security (GenoPri'15), San Jose (2015)
76. Duverle, D., Kawasaki, S., Yamada, Y., et al.: Privacy-preserving statistical analysis by exact logistic regression. In: 2nd International Workshop on Genome Privacy and Security (GenoPri'15), San Jose (2015)
77. Kantarciooglu, M., Jiang, W., Liu, Y., et al.: A cryptographic approach to securely share and query genomic sequences. *IEEE Trans. Inf. Technol. Biomed.* **12**, 606–617 (2008). doi:[10.1109/TITB.2007.908465](https://doi.org/10.1109/TITB.2007.908465)
78. Malin, B.A.: Protecting genomic sequence anonymity with generalization lattices. *Methods Inf. Med.* **44**, 687–692 (2005). <http://www.ncbi.nlm.nih.gov/pubmed/16400377>. Accessed 12 Jan 2012
79. Loukides, G., Gkoulalas-Divanis, A., Malin, B.: Anonymization of electronic medical records for validating genome-wide association studies. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 7898–7903 (2010). doi:[10.1073/pnas.0911686107](https://doi.org/10.1073/pnas.0911686107)
80. Yu, F., Rybar, M., Uhler, C., et al.: Differentially-private logistic regression for detecting multiple-SNP association in GWAS databases. In: Domingo-Ferrer, J., (ed.) *Privacy in Statistical Databases*, pp. 170–184. Springer, Cham (2010). doi:[10.1007/978-3-540-87471-3](https://doi.org/10.1007/978-3-540-87471-3)
81. Wang, S., Mohammed, N., Chen, R.: Differentially private genome data dissemination through top-down specialization. *BMC Med. Inform. Decis. Mak.* **14**, S2 (2014). doi:[10.1186/1472-6947-14-S1-S2](https://doi.org/10.1186/1472-6947-14-S1-S2)
82. Johnson, A., Shmatikov, V.: Privacy-preserving data exploration in genome-wide association studies. In: Proceedings of the 19th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining - KDD '13, p. 1079. ACM Press, New York (2013). doi:[10.1145/2487575.2487687](https://doi.org/10.1145/2487575.2487687)
83. Uhler, C., Slavkovic, A.B., Fienberg, S.E.: Privacy-preserving data sharing for genome-wide association studies. *J. Priv. Confidentiality* **5**, 137–166 (2013)
84. Yu, F., Fienberg, S.E., Slavkovic, A.B., et al.: Scalable privacy-preserving data sharing methodology for genome-wide association studies. *J. Biomed. Inform.* **50**, 133–141 (2014). doi:[10.1016/j.jbi.2014.01.008](https://doi.org/10.1016/j.jbi.2014.01.008)
85. Yu, F., Ji, Z.: Scalable privacy-preserving data sharing methodology for genome-wide association studies: an application to iDASH healthcare privacy protection challenge. *BMC Med. Inform. Decis. Mak.* **14**, S3 (2014). doi:[10.1186/1472-6947-14-S1-S3](https://doi.org/10.1186/1472-6947-14-S1-S3)
86. De Cristofaro, E.: Genomic privacy and the rise of a new research community. *IEEE Secur. Priv.* **12**, 80–83 (2014). doi:[10.1109/MSP.2014.24](https://doi.org/10.1109/MSP.2014.24)
87. 2nd International Workshop on Genome Privacy and Security (GenoPri 2015). <http://www.genopri.org/>. Accessed 30 Mar 2015
88. Ohno-Machado, L., Bafna, V., Boxwala, A.A., et al.: iDASH: integrating data for analysis, anonymization, and sharing. *J. Am. Med. Inform. Assoc.* **19**, 196–201 (2012)
89. Jiang, X., Zhao, Y., Wang, X., et al.: A community assessment of privacy preserving techniques for human genomes. *BMC Med. Inform. Decis. Mak.* **14**(Suppl 1), S1 (2014). doi:[10.1186/1472-6947-14-S1-S1](https://doi.org/10.1186/1472-6947-14-S1-S1)

Chapter 17

Private Genome Data Dissemination

Noman Mohammed, Shuang Wang, Rui Chen, and Xiaoqian Jiang

Abstract With the rapid advances in genome sequencing technology, the collection and analysis of genome data have been made easier than ever before. In this course, sharing genome data plays a key role in enabling and facilitating significant medical breakthroughs. However, substantial privacy concerns have been raised on genome data dissemination. Such concerns are further exacerbated by several recently discovered privacy attacks. In this chapter, we review some of these privacy attacks on genome data and the current practices for privacy protection. We discuss the existing work on privacy protection strategies for genome data. We also introduce a very recent effort to disseminating genome data while satisfying differential privacy, a rigorous privacy model that is widely adopted for privacy protection. The proposed algorithm splits raw genome sequences into blocks, subdivides the blocks in a top-down fashion, and finally adds noise to counts in order to preserve privacy. It has been empirically shown that it can retain essential data utility to support different genome data analysis tasks.

17.1 Introduction

In the past decades, genome sequencing technology has experienced unprecedented development. The Human Genome Project (HGP), initiated in 1990, took 13 years to complete at a total cost of \$3 billion, while nowadays it takes only 2 or 3 days for a whole genome sequence at the cost of \$6K [1]. The readily availability of genome data has spawned many new exciting research areas, ranging from understanding the mechanisms of cellular functions to identifying criminals. There has been no

N. Mohammed (✉)

Department of Computer Science, University of Manitoba, Winnipeg, MB R3T 2N2, Canada
e-mail: noman@cs.umanitoba.ca

S. Wang • X. Jiang

Department of Biomedical Informatics, University of California, San Diego, La Jolla, CA, USA
e-mail: shw070@ucsd.edu; x1jiang@ucsd.edu

R. Chen

Samsung Research America, Mountain View, CA, USA
e-mail: rui.chen1@samsung.com

doubt that genome data analysis generates interesting scientific discoveries and enables significant medical breakthroughs. However, substantial privacy concerns have been raised on the dissemination of genome data. The public has realized that a personally identifiable genomic segment already gives an adversary access to a wealth of information about the individual and his or her genetic relatives [2], exposing their privacy to considerable risks.

Genomic data leakage has serious implications for research participants such as discrimination for employment, insurance, or education [1]. Recent research results show that given some background information about an individual, an adversary can identify or learn sensitive information about the victim from the de-identified data. For example, Homer's attack [3] demonstrated that it is possible to identify a genome-wide association study (GWAS) participant from the allele frequencies of a large number of single-nucleotide polymorphisms (SNPs). As a consequence, the U.S. National Institutes of Health (NIH) has forbidden public access to most aggregate research results to protect privacy. Later, Wang et al. [4] showed an even higher risk that individuals could be actually identified from a relatively small set of statistics, such as those routinely published in GWAS papers. There are also many other attacks revealed recently [5–7], which could result in harm to the privacy of individuals. Therefore, there has been a growing demand to promote privacy protection for genome data dissemination.

In the United States, the Health Insurance Portability and Accountability Act (HIPAA) [8] establishes the Privacy Rule to protect health information. The Privacy Rule defines an operational approach, called Safe Harbor that removes 18 HIPAA-specified identifiers to achieve some degree of “de-identification”. Since genome data are biometrics, it would be natural to remove these data from “de-identified” data sets. However, there is no explicit clarification of de-identified genomic data by the Institute of Medicine (IOM) or HIPAA regulations.

There have been long and vigorous debates [9, 10] about the current privacy rules for Human Genomic Studies (HGS). Some researchers contend that existing privacy rules are not adequate for the protection of genomic information [4, 11], as the technological evolution and the increasing accessibility of data cause the “de-identified” genome data to be re-identifiable. Others complain that privacy regulations impede effective data access and use for research [9, 12], as genomic data are most useful when presented in high quality, sufficient samples, and associated with an individual's medical history, etc. Recently, the Presidential Commission for the Study of Bioethical Issues published a report about privacy and progress in Whole Genome Sequencing (WGS) [12]. The report concludes that under current privacy rules, genome privacy is not adequately protected and that at the same time genomic researchers and data owners cannot effectively access and share data. To address these limitations, there have been some pioneering efforts on developing practical privacy-preserving technology solutions to genome data sharing.

In this chapter, we present an approach to disseminate genomic data in a privacy-preserving manner. The privacy guarantee is guarded by the rigorous *differential privacy* model [13]. Differential privacy is a rigorous privacy model that makes no

assumption about an adversary's background knowledge. A differentially-private mechanism ensures that the probability of any output (released data) is equally likely from all nearly identical input datasets and thus guarantees that all outputs are insensitive to any individual's data. In other words, an individual's privacy is not at risk because of his or her participation in the dataset. The proposed approach uses a top-down structure to split long sequences into segments before adding noise to mask record owners' identity, which demonstrates promising utility with a desirable computational complexity.

The rest of the chapter is organized as follows. An overview of the related work is presented in Sect. 17.2. Section 17.3 describes the problem statement and details the data privacy requirement and the utility criteria. Section 17.4 describes the proposed algorithm and analyzes the privacy guarantee and the computational cost of the proposed method. Experimental results are presented in Sect. 17.5, and finally, Sect. 17.6 concludes the chapter.

17.2 Literature Review

17.2.1 Privacy Attacks and Current Practices

In general, the traditional consent mechanism that allows data disclosure without de-identification is not suitable for genomic data [14]. It is often impossible to obtain consent for an unknown future research study without creating any bias in the sample. In addition, the consent mechanism also raises a number of ethical issues (e.g., withdrawal from future research, promise of proper de-identification, etc.) [15].

A common de-identification technique is to remove all explicit identifying attributes (e.g., name and SSN) and to assign each record a random number prior to data sharing. However, this technique is vulnerable to identity disclosure attacks as demonstrated by a recent study that successfully identified the participants of the Personal Genome Project (PGP) through public demographic data [16].

Given that a genome sequence is a strong personal identifier, several organizations including the U.S. National Institutes of Health (NIH) initially adopted a two-tier access model: controlled access and open access [17]. Individual-level biomedical data are only available to approved researchers (i.e., data requesting institutions) based on proper agreements, while summary statistics, which are useful for meta-GWAS analyses, can be disclosed publicly. Unfortunately, some recent studies have shown that from summary statistics adversaries can already learn sensitive information [3, 4]. This is known as an attribute disclosure attack. Currently, there are a number of techniques for breaching biomedical data privacy [18]. In response, the NIH and other data custodians have moved summary statistics from open access to controlled access. For example, the Database of Genotypes and Phenotypes (dbGaP) [19], which is a popular public database recommended by

Genome Canada for sharing biomedical data, no longer provides open access to summary statistics. Although such policy provides enhanced privacy protection, it largely limits researchers' ability to conduct timely research.

In the United States, the Health Insurance Portability and Accountability Act (HIPAA) provides two standards for health data sharing without requiring patients' consent. The Safe Harbor Standard, which is also used by health organizations in Canada [20], considers a dataset properly de-identified if the 18 specific attributes are removed. However, genomic data along with other types of data (e.g., diagnostic code) are not part of these specified attributes [21]. The alternative Statistical Standard requires an expert to certify that the risk of re-identification is very small from the disclosed data. However, there is no specific guideline to address how to achieve Statistical Standard for genomic data.

The current practice of genomic data sharing is based on the controlled access model, and privacy is ensured through a data-use certification (DUC) agreement. However, an agreement cannot prevent an insider from intentionally performing privacy attacks or even stealing data. A recent article reported several violations of such agreements (e.g., investigators sharing controlled-access data with unapproved parties) that occurred in the last 6-year period [22].

17.2.2 Privacy Preserving Techniques

Privacy-preserving data sharing techniques study how to transform raw data into a version that is immune to privacy attacks but that still preserves useful information for data analysis. Existing techniques are primarily based on two major privacy models: k -anonymity [23] and differential privacy [13]. Data sharing techniques adopting the k -anonymity model require that no individual should be identifiable from a group of size smaller than k based on the values of quasi-identifiers (e.g., age, gender, date of birth). In spite of its wide application in the healthcare domain [21, 24], recent research results indicate that k -anonymity based techniques are vulnerable to an adversary's background knowledge [25–28]. This has stimulated a discussion in the research community in favor of the differential privacy model, which provides provable privacy guarantees independent of an adversary's background knowledge.

To satisfy a specific privacy model, certain anonymization techniques have to be developed to achieve a reasonable trade-off between privacy and utility. While many anonymization techniques have been proposed for various types of data (i.e., relational [29, 30], set-valued [31], spatio-temporal data [32]), the problem of genomic data anonymization has been little studied. Recent methods [33–35] propose to generalize genomic data to achieve the k -anonymity privacy model. Malin [35] presents how to anonymize genome sequences. Loukides et al. [33] and Heatherly et al. [34] propose to anonymize only health data (i.e., no protection for genome sequences). However, as mentioned before, all methods adopting k -anonymity as the underlying privacy model are vulnerable to the recently discovered

privacy attacks [25–28]. More recently, differentially private mechanisms [36–38] have been proposed for genomic data. However, these techniques only release some aggregate information or target specific data analysis tasks (e.g., minor allele frequencies, top- k most relevant SNPs). They do not support individual-level data sharing, which could be of much greater interest to the research community.

As another possible direction, cryptography-based methods have also been suggested for distributed genomic data sharing [39, 40]. Data custodians outsource their data securely through homomorphic encryption to a third party that carries out computations on the encrypted data. However, these techniques have several shortcomings. First, they assume the existence of a trusted third party [39] or tamper-resistant hardware [40]. These assumptions may not be practical in most real-life applications. Second, these techniques only support a limited set of aggregate queries and do not enable to share individual-level data. Finally, these techniques suffer from one main drawback: the aggregate output has no privacy guarantee. In the rest of this chapter, we will introduce a novel differentially private data dissemination technique that supports individual-level genome data sharing.

17.3 Problem Statement

Suppose a data owner has a data table $D(A^i, A^{snp})$ and wants to release an anonymous data table \hat{D} to the public for data analysis. The attributes in D are classified into two categories: (1) An *explicit identifier* attribute that explicitly identifies an individual, such as *SSN*, and *Name*. These attributes are removed before releasing the data as per the HIPAA Privacy Rule [41]; (2) A multi-set of SNPs (genomic data), which is denoted by A^{snp} , for each individual in the data table D . For example in Table 17.1, A^i and A^{snp} are *ID* and *Genomic data* attributes, respectively.

Given a data table D , our objective is to generate an anonymized data table \hat{D} such that \hat{D} satisfies ϵ -differential privacy, and preserves as much utility as possible for data analysis. Next, we introduce differential privacy and data utility models.

Table 17.1 Raw genome data

ID	Genomic data
1	AG CC CC GG CT GG AA CC
2	AG CC CC GG TT GG AA CC
3	AA CC CC GG TT GG AA CC
4	AG CT CT AG CT AG AG CT
5	GG CT CT AG CC GG AA CC
6	AA CC CC GG TT GG AA CC
7	AG CT CT AG CT AG AG CT
8	AA CC CC GG TT GG AA CC
9	GG CT TT AG CC AG AA CC
10	AG CT CT GG CT AG AA CC

17.3.1 Privacy Protection Model

Differential privacy is a recent privacy definition that provides a strong privacy guarantee. It guarantees that an adversary (even with arbitrary background knowledge) learns nothing more about an individual from the released data set, regardless of whether her record is present or absent in the original data. Informally, differential privacy requires that any computational output be insensitive to any particular record. Therefore, from an individual's point of view, the output is computed as if from a data set that does not contain his or her record. Formally, differential privacy is defined as follows.

Definition 17.1 (ϵ -Differential Privacy [13]). A randomized algorithm Ag is differentially private if for all data sets D and D' whose symmetric difference is at most one record (i.e., $|D \Delta D'| \leq 1$), and for all possible anonymized data sets \hat{D} ,

$$\Pr[Ag(D) = \hat{D}] \leq e^\epsilon \times \Pr[Ag(D') = \hat{D}], \quad (17.1)$$

where the probability is taken over the randomness of Ag .

The privacy level is controlled by the parameter ϵ . A smaller value of ϵ provides a strong privacy guarantee. A standard mechanism to achieve differential privacy is the Laplace mechanism [13]. Its key idea is to add properly calibrated Laplace noise to the true output of a function in order to mask the impact of any single record. The maximal impact of a record to a function f 's output is called its *sensitivity*.

Definition 17.2 (Sensitivity [13]). For a function $f : D \rightarrow \mathbb{R}^d$, the sensitivity of f is $\Delta f = \max_{D,D'} \|f(D) - f(D')\|_1$ for all databases such that $|D \Delta D'| \leq 1$.

Given a function's sensitivity and a privacy parameter, the Laplace mechanism is given as follows.

Theorem 17.1 ([13]). For any function $f : D \rightarrow \mathbb{R}^d$, the mechanism Ag ,

$$Ag(D) = f(D) + \langle \text{Lap}_1\left(\frac{\Delta f}{\epsilon}\right), \dots, \text{Lap}_d\left(\frac{\Delta f}{\epsilon}\right) \rangle \quad (17.2)$$

gives ϵ -differential privacy, where $\text{Lap}_i\left(\frac{\Delta f}{\epsilon}\right)$ are i.i.d Laplace variables with scale parameter $\frac{\Delta f}{\epsilon}$.

17.3.2 Privacy Attack Model

The likelihood ratio test [42] provides an upper bound on the power of any method for the detection of an individual in a cohort, using the following formula:

$$\bar{L} = \sum_j^m \left(x_j \log \frac{\hat{p}_j}{p_j} + (1 - x_j) \log \frac{1 - \hat{p}_j}{1 - p_j} \right),$$

where x_j is either 0 (i.e., major allele) or 1 (i.e., minor allele), m is the number of SNPs, p_j is the allele frequency of SNP j in the population and \hat{p}_j is that in a pool.

A statistic \bar{L} measure the probability that a subject in the case group will be re-identified. The re-identification risk is considered to be high if the LR test statistic of individual's SNVs is significantly greater than those of individuals who are not in the same group.

17.3.3 Utility Criteria

We use a case-control association χ^2 test to evaluate the utility of a differentially private data. The test has the following form: $\chi^2 = \sum_i^r \sum_j^c \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$, where r is the number of rows, c is the number of columns, $O_{i,j}$ is observed frequencies, and $E_{i,j}$ is expected frequencies. The χ^2 test statistic provide a measure of how close the observed frequencies are to the expected frequencies. Suppose that the observed allele counts (e.g., for allele "A" and "T") for the case group are a and b , respectively, for total of $r = a + b$. Similarly, we define the observed counts for the same allele (i.e., "A" and "T") in the control group as c and d , respectively, for total of $s = c + d$. Then, the expected allele frequencies in the case group can be expressed as $(a + c)r/(r + s)$ and $(b + d)r/(r + s)$. The expected allele frequencies in the control group can be expressed as $(a + c)s/(r + s)$ and $(b + d)s/(r + s)$, which measure the expected allele frequencies in both case and control populations.

17.4 Genomic Data Anonymization

In this section, we first present our genomic data anonymization algorithm as described in Algorithm 17.1 and prove that the algorithm is ϵ -differentially private. We then analyze the runtime complexity of the algorithm.

17.4.1 Anonymization Algorithm

The proposed algorithm first divides the genomic data into blocks and then generalizes each block. Thus, the algorithm divides the raw data into several equivalence groups, where all the records within an equivalence group have the same block values. Finally, the algorithm publishes the noisy counts of the groups. Next we elaborate each line of the algorithm.

Table 17.2 Genome data partitioned into blocks

ID	Genomic data			
	Block 1	Block 2	Block 3	Block 4
1	AG CC	CC GG	CT GG	AA CC
2	AG CC	CC GG	TT GG	AA CC
3	AA CC	CC GG	TT GG	AA CC
4	AG CT	CT AG	CT AG	AG CT
5	GG CT	CT AG	CC GG	AA CC
6	AA CC	CC GG	TT GG	AA CC
7	AG CT	CT AG	CT AG	AG CT
8	AA CC	CC GG	TT GG	AA CC
9	GG CT	TT AG	CC AG	AA CC
10	AG CT	CT GG	CT AG	AA CC

Dividing the Raw Data (Line 1) Algorithm 17.1 first divides the raw genomic data into multiple blocks. Each block consists of a number of SNPs. For example, the raw genomic data of Table 17.1 can be divided into four blocks as shown in Table 17.2, where each block consists of two SNPs. These blocks are treated like different attributes and thus enable the proposed algorithm to anonymize high-dimensional genomic data effectively. We denote each block by A_i^{snp} and thus $A^{snp} = \cup A_i^{snp}$.

Note that the sizes of all the blocks do not need to be equal. For example, if there were nine SNPs in Table 17.1 instead of 8, it would be impossible to have all blocks of size two. In such a case, the last block can be bigger than the other blocks. In principle, each block may have a different size, and the proposed algorithm can handle such a scenario.

We do not use any heuristic to determine the size of each block. Block size is always constant, and hence, this step does not use any privacy budget (See Sect. 17.4.2). Experimental results suggest that six SNPs per block yield good result. However, this number may vary depending on the data set in question. It is an interesting research problem to design a heuristic that can determine the optimal size of each block so as to maximize the data utility for a given data set.

Algorithm 17.1 Genomic data anonymization algorithm.

- **Input:** Raw data set D , privacy budget ϵ , and number of specializations h

- **Output:** Anonymized genomic data set \hat{D}

- 1: Divide the genome data into blocks;
 - 2: Generate the taxonomy tree for each block;
 - 3: Initialize every block in D to the topmost value;
 - 4: Initialize Cut_i to include the topmost value;
 - 5: **for** $i = 1$ to h **do**
 - 6: Select $v \in UCut_i$ randomly;
 - 7: Specialize v on D and update $UCut_i$;
 - 8: **return** each leaf node with noisy count ($C + \text{Lap}(1/\epsilon)$)
-

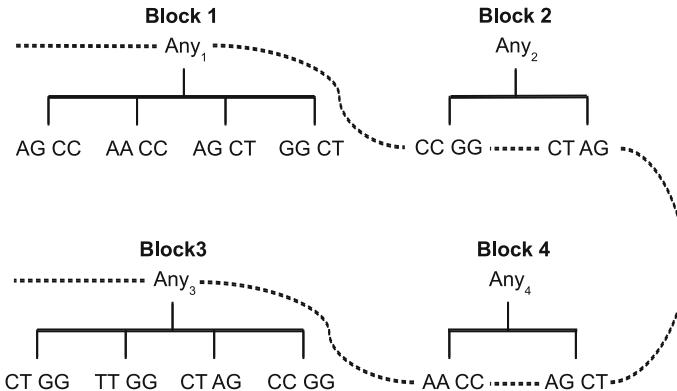


Fig. 17.1 Taxonomy tree of blocks

Generating the Taxonomy Tree (Line 2) A taxonomy tree of a block A_i^{snp} specifies the hierarchy among the values. Figure 17.1 presents the taxonomy trees of Blocks 1–4 (ignore the dashed curve for now) in Table 17.2. A *cut* of the taxonomy tree for a block A_i^{snp} , denoted by Cut_i , contains exactly one value on each root-to-leaf path (more discussion follows).

Ideally, the data owner should provide a taxonomy tree for each block as the knowledge of the taxonomy tree is domain specific. However, if no taxonomy tree is provided, Algorithm 17.1 can generate it by scanning the data set once for each block. For each unique value that appears in the data set, a leaf node is created from the root node Any_1 . For example, four unique values (i.e., AG CC, AA CC, AG CT, and GG CT) appear in Table 17.2 for Block 1; therefore, the corresponding taxonomy tree also has four leaves as shown in Fig. 17.1.

All the generated taxonomy trees have only two levels (i.e., the root and the leaf nodes). However, a data owner can define a multilevel taxonomy tree for each block [43]. A multilevel taxonomy tree provides more flexibility and may preserve more data utility; further investigation is needed in order to validate the benefit of multilevel taxonomy trees.

Data Anonymization (Lines 3–8) The data anonymization starts by creating a single root partition by generalizing all values in $\cup A_i^{snp}$ to the top-most value in their taxonomy trees (Line 3). The initial Cut_i contains the topmost value for each block A_i^{snp} (Line 4).

The specialization starts from the topmost cut and pushes down the cut iteratively by specializing some value in the current cut. The general idea is to anonymize the raw data by a sequence of specializations, starting from the topmost general state as shown in Fig. 17.2. A *specialization*, denoted by $v \rightarrow child(v)$, where $child(v)$ is the set of child values of v , replaces the parent value v with a child value. The specialization process can be viewed as pushing the “cut” of each taxonomy

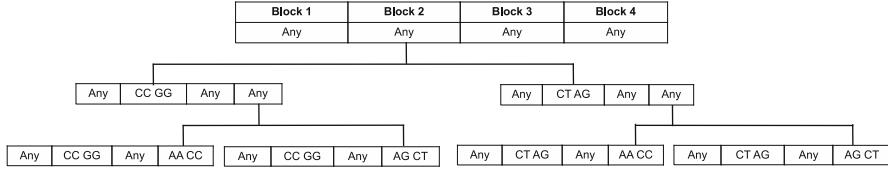


Fig. 17.2 Tree for partitioning records

Table 17.3 Anonymous data
($\epsilon = 1, h = 2$)

Genomic data	Noisy count
Any CC GG Any AA CC	3
Any CC GG Any AG CT	2
Any CT AG Any AA CC	1
Any CT AG Any AG CT	3

tree downwards. Figure 17.1 shows a solution cut indicated by the dashed curve corresponding to the anonymous Table 17.3.

At each iteration, Algorithm 17.1 randomly selects a candidate $v \in \cup Cut_i$ for specialization (Line 6). Candidates can be selected based on their score values, and different heuristics (e.g., information gain) can be used to determine candidates' scores. In future work, we will investigate how to design a scoring function tailored to a specific data utility requirement.

Then, the algorithm specializes v and updates $\cup Cut_i$ (Line 7). Algorithm 17.1 specializes v by recursively distributing the records from the parent partition into disjoint child partitions with more specific values based on the taxonomy tree. The algorithm terminates after a given number of specializations.

Example 17.1. Consider Table 17.1 with $\epsilon = 1$ and $h = 2$, where ϵ is the privacy budget and h is the number of specializations. Initially the algorithm creates one root partition containing all the records that are generalized to $\{Any_1, Any_2, Any_3, Any_4\}$. $\cup Cut_i$ includes $\{Any_1, Any_2, Any_3, Any_4\}$. Let the first specialization be $Any_2 \rightarrow \{CC GG, CT AG\}$. The algorithm creates two new partitions under the root, as shown in Fig. 17.2, and splits data records between them. $\cup Cut_i$ is updated to $\{Any_1, Any_3, Any_4\}$. Suppose that the next specialization is $Any_4 \rightarrow \{AA CC, AG CT\}$, which creates further specialized partitions, as illustrated in Fig. 17.2. ■

Returning the Noisy Counts (Line 9) Finally, Algorithm 17.1 computes the noisy count of each leaf partition to construct the anonymous data table \hat{D} as shown in Table 17.3. The number of leaf partitions is at least 2^h and the exact number depends on the taxonomy tree of the blocks.

Publishing the true counts of each partition violates differential privacy; therefore, a random variable $Lap(\Delta f / \epsilon)$ is added to the true count of each leaf partition, where $\Delta f = 1$, and the noisy counts are published instead.

17.4.2 Privacy Analysis

We now analyze the privacy implication of each of the above steps and quantify the information leakage in terms of privacy budget.

Line 1 The algorithm divides the raw data into blocks, where the block size is a given constant irrespective of the given data set. Since the block generation process is data independent, this step does not require any privacy budget. However, if a heuristic were used to determine the block size, then a portion of privacy budget should be allocated to satisfy differential privacy.

Line 2 We assume that the data owner provides the taxonomy trees. In such a case, this step incurs no privacy leakage and no privacy budget is consumed as the taxonomy trees are generated from public knowledge that is independent of any particular data set.

On the other hand, the alternative approach that we outlined, for a scenario when the taxonomy trees are not provided, needs additional treatment to satisfy differential privacy. It is because, for a different data set \hat{D} , a taxonomy tree may have one more or less leaf node. We argue that taxonomy trees represent the domain knowledge, and therefore, should be part of public information.

Lines 3–8 The algorithm selects a candidate for specialization randomly (Line 7) and iteratively creates child partitions based on the given taxonomy trees (Line 8). Both operations are independent of the underlining data set (the selection process is random and the partitioning process is fixed due to the given taxonomy trees), and therefore no privacy budget is required for the h number of iterations.

Line 9 The algorithm adds Laplace noise $\text{Lap}(1/\epsilon)$ to the true count of each leaf partition and the requisite privacy budget is ϵ due to the *parallel composition property* [44]. The Parallel composition property guarantees that if a sequence of computations are conducted on *disjoint* data sets, then the privacy cost does not accumulate but depends only on the worst guarantee of all the computations. Since the leaf partitions are disjoint (i.e., a record can fall into exactly one leaf partition), the total privacy cost (i.e., the budget required) for this step is ϵ .

In conclusion, Line 1, Line 2, Lines 3–8, and Line 9 use 0, 0, 0, and ϵ privacy budgets, respectively. According to the *sequential composition property* of differential privacy [44], any sequence of computations that each provides differential privacy in isolation also provides differential privacy in sequence. Therefore, Algorithm 17.1 satisfies ϵ -differential privacy.

17.4.3 Computational Complexity

The proposed algorithm is scalable and the runtime is linear to the size of the data set. This is an important property to achieve in the age of big data. In this section, we present a brief analysis of the computational complexity of Algorithm 17.1.

Line 1 Algorithm 17.1 generates the blocks from the raw data. This can be done by scanning the data set once. Thus, the runtime of this step is $O(|D| \times m)$, where $|D|$ is the number of records and m is the number of SNPs.

Line 2 In case, Algorithm 17.1 can also generate the taxonomy trees (if not given) by scanning the data set once. This is can be achieved simultaneously with the previous step (Line 1); hence, there is no additional cost for generating taxonomy trees. Therefore, if there are d/n number of blocks, where n is the block size, then the runtime of this step is $O(|D| \times \frac{d}{n})$.

Lines 3–8 Candidates are selected randomly in each iteration, which requires constant $O(1)$ time (Line 6).

To perform a specialization $v \rightarrow child(v)$, we need to retrieve D_v , the set of data records generalized to v . To facilitate this operation we organize the records in a tree structure as shown in Fig. 17.2. Each leaf partition (node) stores the set of data records having the same generalized block values. This will allow us to calculate the noisy counts in Line 9.

Initially, the tree has only one leaf partition containing all data records, generalized to the topmost value on every block. In each iteration we perform a specialization by refining the leaf partitions and splitting the records among the new child partitions. This operation also requires scanning all the records once per iteration. Thus, the runtime of this step is $O(|D| \times h)$. The value of h is constant and usually very small (around 10), and therefore, can be ignored.

Line 9 The cost of adding Laplace noise is proportional to the number of leaf nodes, which is 2^h . For a small value of h , the number of leaf nodes is insignificant with respect to the size of the data set $|D|$. We therefore can ignore the cost of this step. Note that, we can easily determine the true count of a leaf partition as it keeps track of the set of data records it represents.

Hence, the total runtime of the algorithm is: $O(|D| \times m + |D|) = O(|D| \times m)$.

17.5 Experimental Results

The goal of the proposed framework is to generate differentially private data that can mitigate the attack of likelihood ratio tests, while preserving highly significant SNPs as much as possible. Two data sets (i.e., chr2 and chr10) with 200 participants in case, control and test groups were used in our experiments. The 200 cases are from Personal Genome Project (PGP: <http://www.personalgenomes.org/>), missing values filled by using fastPHASE. The 200 Controls are simulated based on the haplotypes of 174 individuals from CEU population of International HapMap Project (<http://hapmap.ncbi.nlm.nih.gov/>). Besides, the chr2 and chr10 data sets contain 311 SNPs and 610 SNPs, respectively.

Table 17.4 Data utility of chr2 data set with privacy budget of 1.0 and power of 0.01

Cutoff p-value	Accuracy	Sensitivity	Precision	F1-score	# of significant SNPs
5E-02	0.178	1.000	0.079	0.147	22
1E-02	0.211	0.999	0.075	0.140	20
1E-03	0.250	0.948	0.072	0.134	19
1E-05	0.297	1.000	0.060	0.114	14

Table 17.5 Data utility of chr10 data set with privacy budget of 1.0 and power of 0.09

Cutoff p-value	Accuracy	Sensitivity	Precision	F1-score	# of significant SNPs
5E-02	0.301	0.956	0.092	0.168	45
1E-02	0.317	0.903	0.048	0.091	23
1E-03	0.431	1.000	0.041	0.080	15
1E-05	0.577	1.000	0.030	0.058	8

The number of specializations used in our experiment was 5. SNP data were split evenly into $N/6$ blocks, where N is the number of SNP. All the results are based on the average of 100 trials.

Tables 17.4 and 17.5 illustrate the results of the proposed method on chr2 and chr10 data sets with privacy budget of 1.0, where power indicates the ratio of identifiable individuals using the likelihood ratio test in the case group. The power serves as a measurement of the remaining privacy risk in the differentially private results. In Tables 17.4 and 17.5, cutoff p-value thresholds of 5E-2, 1E-2, 1E-3, 1E-5 were used in our experiment, for which four measurements (accuracy, sensitivity, precision and F1-score) were calculated under each method. The last column corresponds to the number of significant SNPs discovered in the original data without adding noise. We can see that the proposed results showed high sensitivities but low precisions on both data sets, which means our method can correctly preserve most true significant SNPs, but with a large amount of false positive reports. Because most SNPs of interest can pass the same filter (e.g., p-value) on both the original data and our differentially private outputs, the latter can serve as a proxy (for exploratory analysis) of the former without losing too much critical information.

Figure 17.3 show the box plots of the data utility in terms of sensitivity and precision for both testing data sets with privacy budget of 1.0 under different cutoff p-values. We can see that the proposed method achieved high sensitivity on both data sets for all cutoff p-values. Moreover, Fig. 17.3 also depict that the precision decreases as the cutoff p-value decreases. Comparing the experiments, results on chr2 are less sensitive in precision than results on chr10 when p-values changes.

Figure 17.4 present the test statistics calculated on case and test groups (i.e., individuals unrelated to both case and control) for both chr2 and chr10 data sets. An individual in the case group can be re-identified with a high confidence if the test statistic obtained from his/her SNP sequence is significantly higher than these of the test group using likelihood ratio test [42]. Figure 17.4 depict that 2 and 18 case

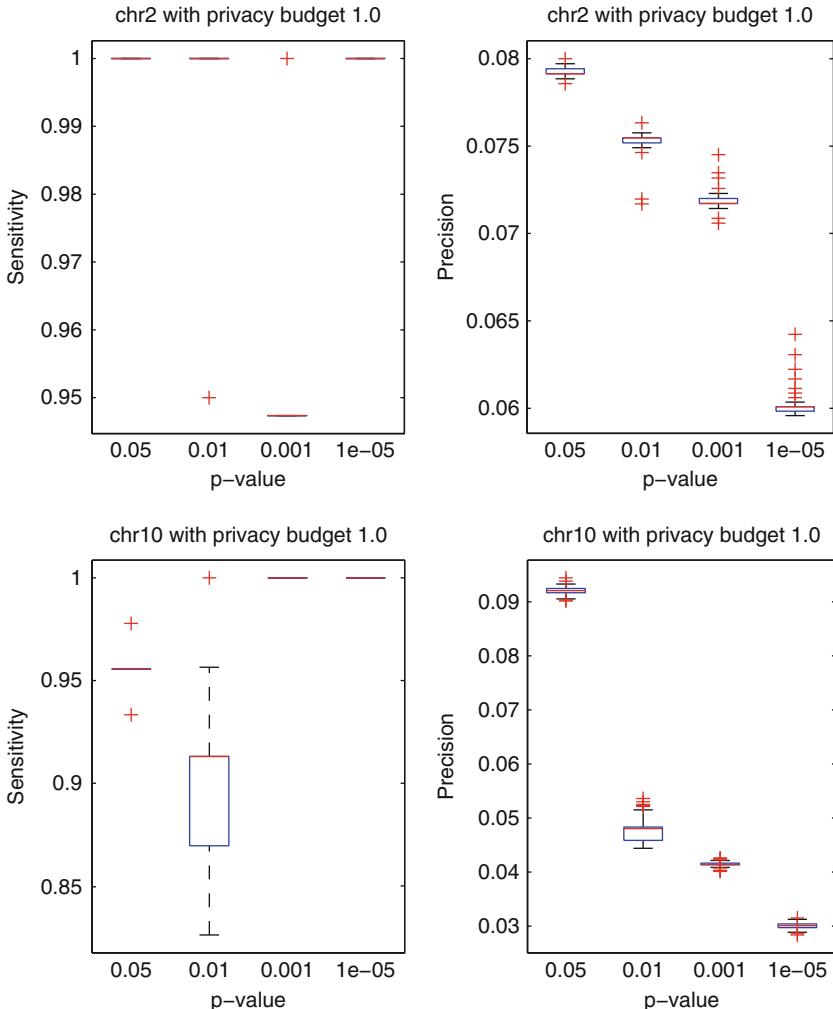


Fig. 17.3 Boxplots of data utility of chr2 and chr10 data with different p-values

individuals have higher test statistic values than 95 % test individuals (i.e., a 5 % false positive rate) in both data sets. The results suggest that the proposed method provides a better privacy protection on a small data set (i.e., chr2 data set) under the same privacy budget.

Finally, Fig. 17.5 show both utility and privacy risk for chr2 and chr10 data sets. To measure the utility, we set the cutoff threshold at 0.1 and measure the sensitivity. By changing privacy budget from 0.1 to 1, we observed no performance gain of sensitivity nor much privacy risk change on chr2 data set, as shown in Fig. 17.5. This seems to indicate that privacy budget 0.1 is sufficient to provide enough protection without destroying the utility.

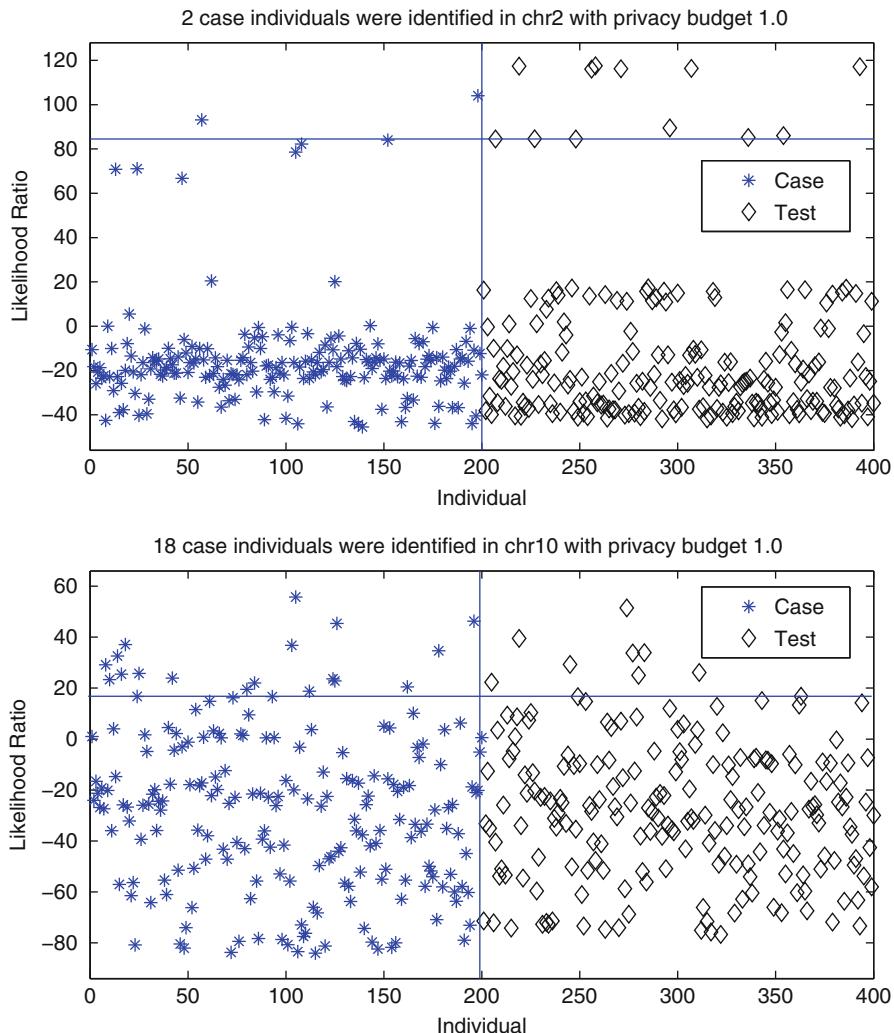


Fig. 17.4 Privacy risk of chr2 and chr10 data. The *star* and *diamond* markers represent the test value of a specific individual in the case (*left*) or test (*right*) group, respectively. The *horizontal line* indicates the 0.95 confidence level for identifying case individuals that are estimated based on the test statistic values of test individuals

We also tested the proposed algorithm on a larger data set (i.e., chr10). Figure 17.5 shows that the proposed algorithm achieves the best sensitivity and the highest number of re-identification risk with privacy budget of 1.0. There is a non-negligible difference in terms of re-identification risk when the privacy budget changes from 0.1 to 0.5 but the difference is not obvious between privacy budgets 0.5 and 1. This indicates that a larger dataset like chr10 needs more privacy budget to protect the privacy of its entities.

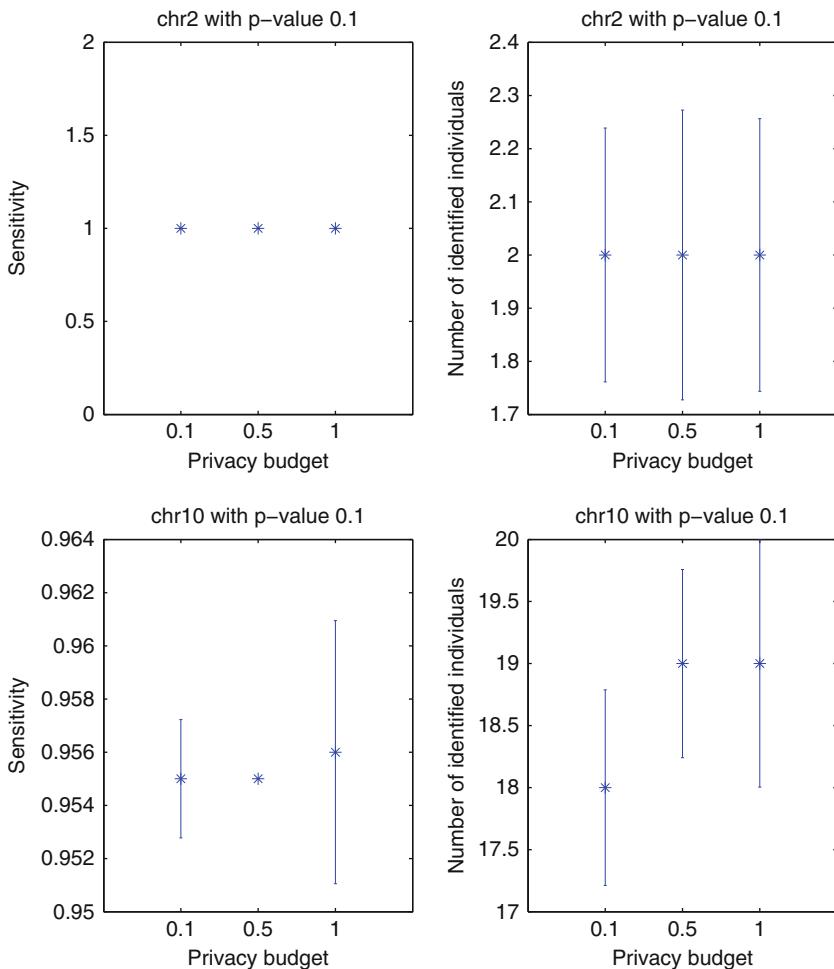


Fig. 17.5 Comparison of data utility and privacy risk for chr2 and chr10 data with different privacy budget

17.6 Conclusion

As an important type of modern medical data, genome data has been incorporated into diverse research disciplines and real-life applications. More and more patient data networks are incorporating or prepare to incorporate genome data along with clinical information to support precision medicine. However, its personal identifying nature has aroused much public concern about the privacy implications of its dissemination, which has been increasingly confirmed by several recently discovered privacy attacks. The current practice of genomic data sharing heavily relies on the controlled access model. Despite the restriction the model poses, it still

cannot provide guaranteed privacy protection. For example, side channel leakage remains to be a big problem to controlled access model.

Advanced technical efforts are indispensable to private genome data dissemination. In this chapter, we gave an overview of the recent developments toward accomplishing this goal. Our focus was on a new differentially private genome data dissemination algorithm. This algorithm supports individual-level data sharing, which is a desideratum for many research areas.

While the existing studies have demonstrated the promise of private genome data sharing, there are still some notable limits. For example, the precision performance of the proposed framework is relatively poor. To make this anonymization process effective, several challenges must be addressed. First, further study is needed to understand how to partition the genome sequences into blocks that are still meaningful. Genomic data is high dimensional and it includes hundreds of Single Nucleotide Polymorphisms (SNPs). Dividing the long sequences into blocks can reduce the high-dimensionality challenge. Second, it might be useful to construct a taxonomy tree for each genome block, which was not tried before. While multilevel taxonomy trees have been proposed for a SNP [35], it is not clear how to construct multilevel trees for a block. Finally, genomic data are shared for different data analysis tasks (e.g., association test, logistic regression); in which case introducing a task specific utility function in the data specialization process may preserve better data utility.

Acknowledgements This article was funded by iDASH (U54HL108460), NHGRI (K99HG 008175), NLM (R00LM011392, R21LM012060), NCBC-linked grant (R01HG007078) and NSERC Discovery Grants (RGPIN-2015-04147).

References

1. Naveed, M., Ayday, E., Clayton, E.W., Fellay, J., Gunter, C.A., Hubaux, J.-P., Malin, B.A., Wang, X.F.: Privacy in the Genomic Era. *ACM Comput. Surv.* to appear
2. Roche, P.A., Annas, G.J.: DNA testing, banking and genetic privacy. *N. Engl. J. Med.* **355**, 545–546 (2006)
3. Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V., Stephan, D.A., Nelson, S.F., Craig, D.W.: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using highdensity SNP genotyping microarrays. *PLoS Genet.* **4**(8), e1000167 (2008)
4. Wang, R., Li, Y.F., Wang, X.F., Tang, H., Zhou, X.: Learning your identity and disease from research papers: information leaks in genome wide association study. In: Proceedings of the 16th ACM Conference on Computer and Communications Security (CCS), New York, pp. 534–544 (2009)
5. Goodrich, M.T.: The mastermind attack on genomic data. In: Proceedings of the 30th IEEE Symposium on Security and Privacy (S&P), pp. 204–218 (2009)
6. Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., Erlich, Y.: Identifying personal genomes by surname inference. *Science* **339**(6117), 321–324 (2013)

7. Rodriguez, L.L., Brooks, L.D., Greenberg, J.H., Green, E.D.: The complexities of genomic identifiability. *Science* **339**(6117), 275–276 (2013)
8. Health Insurance Portability and Accountability Act of 1996. Public L. No. 104–191, 110 Stat. 1936, 1996. <http://www.gpo.gov/fdsys/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf>.
9. Zhou, X., Peng, B., Li, Y., Chen, Y.: To release or not to release: evaluating information leaks in aggregate human-genome data. In: Security ESORICS, Leuven, pp. 1–27 (2011)
10. Weaver, T., Mauret, J., Hayashizaki, Y.: Sharing genomes: an integrated approach to funding, managing and distributing genomic clone resources. *Nat. Rev. Genet.* **5**(11), 861–866 (2004)
11. Malin, B.A., Sweeney, L.A.: How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *J. Biomed. Inform.* **37**(3), 179–192 (2004)
12. Presidential Commission for the Study of Bioethical Issues: Privacy and Progress in Whole Genome Sequencing (October) (2012)
13. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Proceedings of the 3rd Conference on Theory of Cryptography (TCC), pp. 265–284 (2006)
14. Caulfield, T., Knoppers, B.: Consent, privacy and research biobanks: policy brief No. 1. Genomics, Public Policy and Society, Genome Canada (2010)
15. Ogbogu, U., Burningham, S.: Privacy protection and genetic research: where does the public interest lie? *Alberta Law Rev.* **51**(3), 471–496 (2014)
16. Sweeney, L., Abu, A., Winn, J.: Identifying participants in the personal genome project by name (a re-identification experiment) (2013) [arXiv:1304.7605]
17. National Institutes of Health, Modifications to Genome-Wide Association Studies (GWAS) Data Access, 28 August 2008
18. Erlich, Y., Narayanan, A.: Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.* **15**(6), 409–21 (2014)
19. Mailman, M., et al.: The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**(10), 1181–1186 (2007)
20. Emam, K.: Data anonymization practices in clinical research: a descriptive study. Health Canada, Access to Information and Privacy Division (2006).
21. Emam, K.: Methods for the de-identification of electronic health records for genomic research. *Genome Med.* **3**, 25 (2011). doi:10.1186/gm239
22. Paltoo, D., et al.: Data use under the NIH GWAS data sharing policy and future directions. *Nat. Genet.* **46**, 934–938 (2014)
23. Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **10**, 557–570 (2002)
24. Emam, K.: A globally optimal k-anonymity method for the de-identification of health data. *J. Am. Med. Inform. Assoc.* **16**(5), 670–682 (2009)
25. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **1**(1), Article No. 3 (2007)
26. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: a new privacy measure for data publishing. *IEEE Trans. Knowl. Data Eng.* **22**(7), 943–956 (2010)
27. Zhang, L., Jajodia, S., Brodsky, A.: Information disclosure under realistic assumptions: privacy versus optimality. In: Proceedings of the 14th ACM Conference on Computer and Communications Security (CCS), pp. 573–583 (2007)
28. Ganta, S., Kasiviswanathan, S., Smith, A.: Composition attacks and auxiliary information in data privacy. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 265–273 (2008)
29. Fung, B., Wang, K., Yu, P.: Anonymizing classification data for privacy preservation. *IEEE Trans. Knowl. Data Eng.* **19**(5), 711–725 (2007)
30. Mohammed, N., Chen, R., Fung, B.C.M., Yu, P.S.: Differentially private data release for data mining. In: Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 493–501, San Diego, CA (2011)

31. Terrovitis, M., Mamoulis, N., Kalnis, P.: Local and global recoding methods for anonymizing set-valued data. *J. Very Large Data Bases* **20**(1), 83–106 (2011)
32. Fan, L., Xiong, L., Sunderam, V.: Differentially private multi-dimensional time-series release for traffic monitoring. In: Proceedings of the 27th IFIP WG 11.3 Conference on Data and Applications Security and Privacy (2013)
33. Loukides, G., Gkoulalas-Divanis, A., Malin, B.: Anonymization of electronic medical records for validating genome-wide association studies. *Proc. Natl. Acad. Sci. U. S. A.* **107**(17), 7898–7903 (2010)
34. Heatherly, R., Loukides, G., Denny, J., Haines, J., Roden, D., Malin, B.: Enabling genomic-phenomic association discovery without sacrificing anonymity. *PLoS ONE* **8**(2), e53875 (2013)
35. Malin, B.A.: Protecting DNA sequences anonymity with generalization lattices. *Methods Inf. Med.* **12**(1), 687–692 (2005)
36. Johnson, A., Shmatikov, V.: Privacy-preserving data exploration in genome-wide association studies. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 1079–1087 (2013)
37. Chen, R., Peng, Y., Choi, B., Xu, J., Hu, H.: A private DNA motif finding algorithm. *J. Biomed. Inform.* **50**, 122–132 (2014)
38. Yu, F., Fienberg, S.E., Slavkovic, A.B., Uhler, C.: Scalable privacy-preserving data sharing methodology for genome-wide association studies. *J. Biomed. Inform.* **50**, 133–141 (2014)
39. Kantarciooglu, M., Jiang, W., Liu, Y., Malin, B.: A cryptographic approach to securely share and query genomic sequences. *IEEE Trans. Inf. Technol. Biomed.* **12**(5), 606–617 (2008).
40. Canim, M., Kantarciooglu, M., Malin, B.: Secure management of biomedical data with cryptographic hardware. *IEEE Trans. Inf. Technol. Biomed.* **16**(1), 166–175 (2012)
41. Malin, B., Benitez, K., Masys, D.: Never too old for anonymity: a statistical standard for demographic data sharing via the hipaa privacy rule. *J. Am. Med. Inform. Assoc.* **18**(1), 3–10 (2011)
42. Sankararaman, S., Obozinski, G., Jordan, M.I., Halperin, E.: Genomic privacy and limits of individual detection in a pool. *Nat. Genet.* **41**(9), 965–967 (2009)
43. Malin, B.A.: An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. *J. Am. Med. Inform. Assoc.* **12**(1), 28–34 (2005)
44. McSherry, F.: Privacy integrated queries. In: Proceedings of the 35th SIGMOD International Conference on Management of Data (SIGMOD), pp. 19–30 (2009)

Chapter 18

Threats and Solutions for Genomic Data Privacy

Erman Ayday and Jean-Pierre Hubaux

Abstract With the help of rapidly developing technology, DNA sequencing is becoming less expensive. As a consequence, the research in genomics has gained speed in paving the way to personalized (genomic) medicine, and geneticists need large collections of human genomes to further increase this speed. Furthermore, individuals are using their genomes to learn about their (genetic) predispositions to diseases, their ancestries, and even their (genetic) compatibilities with potential partners. This trend has also caused the launch of health-related websites and online social networks (OSNs), in which individuals share their genomic data (e.g., OpenSNP or 23andMe). On the other hand, genomic data carries much sensitive information about its owner. By analyzing the DNA of an individual, it is now possible to learn about his disease predispositions (e.g., for Alzheimer's or Parkinson's), ancestries, and physical attributes. The threat to genomic privacy is magnified by the fact that a person's genome is correlated to his family members' genomes, thus leading to interdependent privacy risks. Thus, in this chapter, focusing on our existing and ongoing work on genomic privacy carried out at EPFL/LCA1, we will first highlight the threats for genomic privacy. Then, we will present the high level descriptions of our solutions to protect the privacy of genomic data and we will discuss future research directions. For a description of the research contributions of other research groups, the reader is referred to Chaps. 16 and 17 of the present volume.

18.1 Threats for Genomic Privacy

Removal of quasi-identifying attributes (e.g., date of birth or zip code) legally protects the privacy of health data. However, it has been shown that anonymization

E. Ayday (✉)

Department of Computer Engineering, Bilkent University, Ankara, Turkey
e-mail: erman@cs.bilkent.edu.tr

J.-P. Hubaux

Institute of Communication Systems, Ecole Polytechnique Fédérale de Lausanne,
Lausanne, Switzerland
e-mail: jean-pierre.hubaux@epfl.ch

is an ineffective technique for genomic data [16, 18, 20]. For example, an adversary can infer the phenotype of the donor of an anonymized genome and use this information to identify the anonymous donor.

For instance, genomic variants on the Y chromosome are correlated with the last name (for males). This last name can be inferred using public genealogy databases. With further effort (e.g., using voter registration forms) the complete identity of the individual can also be revealed [18]. Also, unique features in patient-location visit patterns in a distributed healthcare environment can be used to link the genomic data to the identity of the individuals in publicly available records [33]. Furthermore, it has been shown that Personal Genome Project (PGP) participants can be identified based on their demographics without using any genomic information [42].

The identity of a participant of a genomic study can also be revealed by using a second sample, that is, part of the DNA information from the individual and the results of the corresponding clinical study [9, 16, 21, 25, 43]. For this reason even a small set of variants (e.g., single nucleotide variants - SNPs) of the individual might be sufficient as the second sample. For example, it is shown that as few as 100 SNPs are enough to uniquely distinguish one individual from others [31]. Homer et al. [21] prove that the presence of an individual in a case group can be determined by using aggregate allele frequencies and his DNA profile. Homer's attack demonstrates that it is possible to identify a participant of a Genome-wide association study (GWAS) by analyzing the allele frequencies of a large number of SNPs. Wang et al. [43] showed a higher risk that individuals can actually be identified from a relatively small set of statistics such as those routinely published in GWAS papers. In particular, they show that the presence of an individual in the case group can be determined based upon the pairwise correlation (i.e., linkage disequilibrium) among as few as a couple of hundred SNPs. While the methodology introduced in [21] requires on the order of 10,000 SNPs (of the target individual), this new attack requires only on the order of hundreds. Another similar attack involves the association of DNA sequences to personal names, through diagnosis codes [32].

In another recent study [16], Gitschier shows that a combination of information from genealogical registries and a haplotype analysis of the Y chromosome collected for The HapMap Project, allows for the prediction of the last names of a number of individuals held in the HapMap database. Thus, releasing (aggregate) genomic data is currently banned by many institutions due to this privacy risk. Zhou et al. [45], study the privacy risks of releasing aggregate genomic data. They propose a risk-scale system to classify aggregate data and a guide for their release.

Some believe that they have nothing to hide about their genetic structure, hence they might decide to give full consent for the publication of their genomes on the Internet to help genomic research. However, our DNA sequences are highly correlated to our relatives' sequences. The DNA sequences between two random human beings are more than 99.5 % similar, and this value is even higher for closely related people. Consequently, somebody revealing his genome does not only damage his own genomic privacy, but also puts his relatives' privacy at risk [41].

Moreover, currently, a person does not need consent from his relatives to share his genome online. This is precisely where the interesting part of the story begins: *kin genomic privacy*.

18.1.1 Kin Genomic Privacy

A recent New York Times' article¹ reports the controversy about sequencing and publishing, without the permission of her family, the genome of Henrietta Lacks (who died in 1951). On the one hand, the family members think that her genome is private family information and it should not be published without the consent of the family. On the other hand, some scientists argued that the genomes of current family members have changed so much over time (due to gene mixing during reproduction), that nothing accurate could be told about the genomes of current family members by using Henrietta Lacks' genome. As we have shown in [23] (that we briefly describe hereafter), they were wrong. Minutes after Henrietta Lacks' genome was uploaded to a public website called SNPedia, researchers produced a report full of personal information about Henrietta Lacks. Later, the genome was taken offline, but it had already been downloaded by several people, hence both her and (partially) the Lacks family's genomic privacy was already lost.

Unfortunately, the Lacks, even though possibly the most publicized family facing this problem, are not the only family facing this threat. Genomes of thousands of individuals are available online. Once the identity of a genome donor is known, an attacker can learn about his relatives (or his family tree) by using an auxiliary side channel, such as an online social network (OSN), and infer significant information about the DNA sequences of the donor's relatives. We will show the feasibility of such an attack and evaluate the privacy risks by using publicly available data on the Web.

Although the researchers took Henrietta Lacks' genome offline from SNPedia, other databases continue to publish portions of her genomic data. Publishing only portions of a genome does not, however, completely hide the unpublished portions; even if a person reveals only a part of his genome, other parts can be inferred using the statistical relationships between the nucleotides in his DNA. For example, James Watson, co-discoverer of DNA, made his whole DNA sequence publicly available, with the exception of one gene known as Apolipoprotein E (ApoE), one of the strongest predictors for the development of Alzheimer's disease. However, later it was shown that the correlation (called *linkage disequilibrium* by geneticists) between one or multiple polymorphisms and ApoE can be used to predict the ApoE status [35]. Thus, an attacker can also use these statistical relationships (which are publicly available) to infer the DNA sequences of a donor's family members, even

¹<http://www.nytimes.com/2013/03/24/opinion/sunday/the-immortal-life-of-henrietta-lacks-the-sequel.html?pagewanted=all>

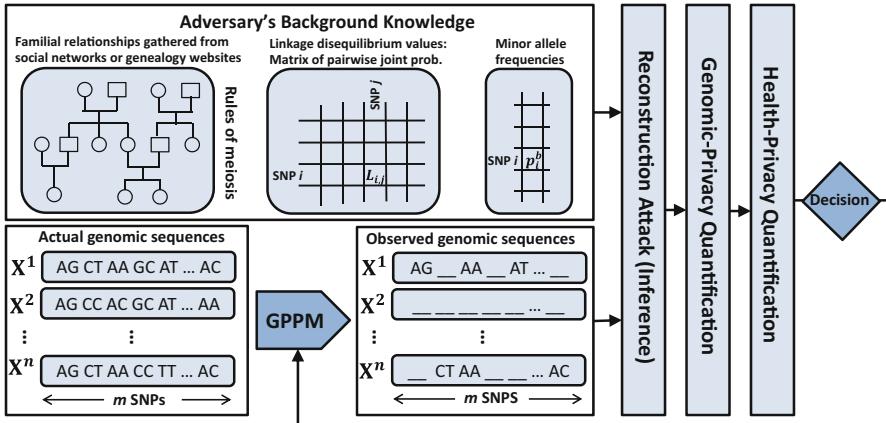


Fig. 18.1 Overview of the proposed framework to quantify kin genomic privacy [23]. Each vector X^i ($i \in \{1, \dots, n\}$) includes the set of SNPs for an individual in the targeted family. Furthermore, each letter pair in X^i represents a SNP x_j^i ; and for simplicity, each SNP x_j^i can be represented using $\{BB, Bb, bb\}$ or $\{0, 1, 2\}$. Once the health privacy is quantified, the family should ideally decide whether to reveal less or more of their genomic information through the genomic-privacy preserving mechanism (GPPM)

if the donor shares only part of his genome. It is important to note that these privacy threats not only jeopardize kin genomic privacy, but, if not properly addressed, these issues could also hamper genomic research due to untimely fear of potential misuse of genomic information.

In [23], we evaluated the genomic privacy of an individual threatened by his relatives revealing their genomes. Focusing on the most common genetic variant in human population, single nucleotide polymorphism (SNP),² and considering the statistical relationships between the SNPs on the DNA sequence, we quantify the loss in genomic privacy of individuals when one or more of their family members' genomes are (either partially or fully) revealed. To achieve this goal, first, we design a reconstruction attack based on a well-known statistical inference technique. The computational complexity of the traditional ways of realizing such inference grows exponentially with the number of SNPs (which is on the order of tens of millions) and relatives. Therefore, in order to infer the values of the unknown SNPs in linear complexity, we represent the SNPs, family relationships and the statistical relationships between SNPs on a factor graph and use the belief propagation algorithm [30, 36] for inference. Then, using various metrics, we quantify the genomic privacy of individuals and show the decrease in their privacy level caused

²A SNP occurs when a nucleotide (at a specific position on the DNA) varies between individuals of a given population. SNPs carry privacy-sensitive information about individuals' health. Recent discoveries show that the susceptibility of an individual to several diseases can be computed from his or her SNPs.

Table 18.1 Frequently used notations

F	Set of family members in the targeted family
S	Set of SNP IDs
x_j^i	Value of SNP j for individual i , $x_j^i \in \{0, 1, 2\}$
\mathbf{X}^i	Set of SNPs for individual i
\mathbb{X}	$n \times m$ matrix that stores the values of the SNPs of all family members
\mathbb{X}_U	Set of SNPs from \mathbb{X} whose values are unknown
\mathbb{X}_K	Set of SNPs from \mathbb{X} whose values are known by the adversary
$\mathcal{F}_R()$	Function representing the Mendelian inheritance probabilities
\mathbb{L}	$m \times m$ matrix representing the pairwise linkage disequilibrium between the SNPs in \mathbf{S}
$\mathbb{L}_{i,j}$	Entry of \mathbb{L} at row i and column j
P	Set of minor allele probabilities (or MAF) of the SNPs in \mathbf{S}

by the published genomes of their family members. We also quantify the health privacy of the individuals by considering their (genetic) predisposition to certain serious diseases. We evaluate the proposed inference attack and show its efficiency and accuracy by using real genomic data of a pedigree.

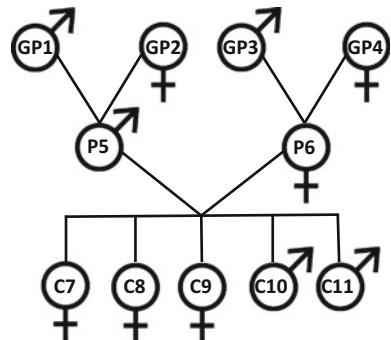
In the following, we formalize our approach and present the different components that will allow us to quantify kin genomic privacy. Figure 18.1 gives an overview of the framework. In order to facilitate future references, frequently used notations are listed in Table 18.1.

In a nutshell, the goal of the adversary is to infer some *targeted SNPs* of a member (or multiple members) of a *targeted family*. We define \mathbf{F} to be the set of family members in the targeted family (whose family tree, showing the familial connections between the members, is denoted as \mathcal{G}_F) and \mathbf{S} to be the set of SNP IDs (i.e., positions on the DNA sequence), where $|\mathbf{F}| = n$ and $|\mathbf{S}| = m$. Note that the SNP IDs are the same for all the members of the family. We also let x_j^i be the value of SNP j ($j \in \mathbf{S}$) for individual i ($i \in \mathbf{F}$), where $x_j^i \in \{0, 1, 2\}$ (a SNP can only be in one of these three states). Furthermore, $\mathbf{X}^i = \{x_j^i : j \in \mathbf{S}, i \in \mathbf{F}\}$ represents the set of SNPs for individual i . We let \mathbb{X} be the $n \times m$ matrix that stores the values of the SNPs of all family members. Some entries of \mathbb{X} might be known by the adversary (the observed genomic data of one or more family members) and others might be unknown. We denote the set of SNPs from \mathbb{X} whose values are unknown as \mathbb{X}_U , and the set of SNPs from \mathbb{X} whose values are known (by the adversary) as \mathbb{X}_K .

$\mathcal{F}_R(x_j^M, x_j^F, x_j^C)$ is the function representing the Mendelian inheritance probabilities, where (M, F, C) represent mother, father, and child, respectively. The $m \times m$ matrix \mathbb{L} represents the pairwise linkage disequilibrium (LD)³ between the SNPs in \mathbf{S} , $\mathbb{L}_{i,j}$ refers to the matrix entry at row i and column j . $\mathbb{L}_{i,j} > 0$ if i and j are in LD, and $\mathbb{L}_{i,j} = 0$ if these two SNPs are independent (i.e., there is no LD between them).

³LD can be thought as a correlation between two variables.

Fig. 18.2 Family tree of *CEPH/Utah Pedigree 1463* consisting of the 11 family members that were considered. The notations GP , P , and C stand for “grandparent”, “parent”, and “child”, respectively. Also, the symbols ♂ and ♀ represent the male and female family members, respectively



$\mathbf{P} = \{p_i^b : i \in \mathbf{S}\}$ represents the set of minor allele probabilities (or MAF) of the SNPs in \mathbf{S} . Finally, note that a joint probability $p(x_i, x_j)$ can be derived from $\mathbb{L}_{i,j}$, p_i^b , and p_j^b .

The adversary carries out a reconstruction attack to infer \mathbb{X}_U by relying on his background knowledge, $\mathcal{F}_R(x_j^M, x_j^F, x_j^C)$, \mathbb{L} , \mathbf{P} , and on his observation \mathbb{X}_K . We formulate the reconstruction attack (on determining the values of the targeted SNPs) as finding the marginal probability distributions of unknown variables \mathbb{X}_U , given the known values in \mathbb{X}_K , familial relationships, and the publicly available statistical information. To run this attack in an efficient way, we formulate the problem on a graphical model (factor graph) and use the belief propagation algorithm for inference. Once the targeted SNPs are inferred by the adversary, we evaluate genomic and health privacy of the family members based on the adversary’s success and his certainty about the targeted SNPs and the diseases they reveal. Finally, we discuss some ideas to preserve the individuals’ genomic and health privacy.

For the evaluation, we used the *CEPH/Utah Pedigree 1463* that contains the partial DNA sequences of 17 family members (4 grandparents, 2 parents, and 11 children) [10]. As shown in Fig. 18.2, we only used the first 5 (out of 11) children (without any particular selection criteria) for our evaluation because (i) 11 is much above the average number of children per family, (ii) we observe that the strength of adversary’s inference does not increase further (due to the children’s revealed genomes) when more than 5 children’s genomes are revealed, and (iii) the belief propagation algorithm might have convergence issues due to the number of loops in the factor graph, and this number increases with the number of children.

We construct \mathbf{S} from 100 SNPs on chromosome 1. Among these 100 SNPs, each SNP is in LD with 5 other SNPs on average. Furthermore, the strength of the LD varies between 0.5 and 1. We note that we only use 100 SNPs for this study as the LD values are not yet completely defined over all SNPs, and the definition of such values is still an ongoing research. We define a target individual from the CEPH family, construct the set \mathbb{X}_U from his/her SNPs, and sequentially reveal other family members’ SNPs (excluding the target individual) to observe the decrease in the genomic privacy of the target individual. We start revealing from the most distant family members to the target individual (in terms of number of hops in Fig. 18.2)

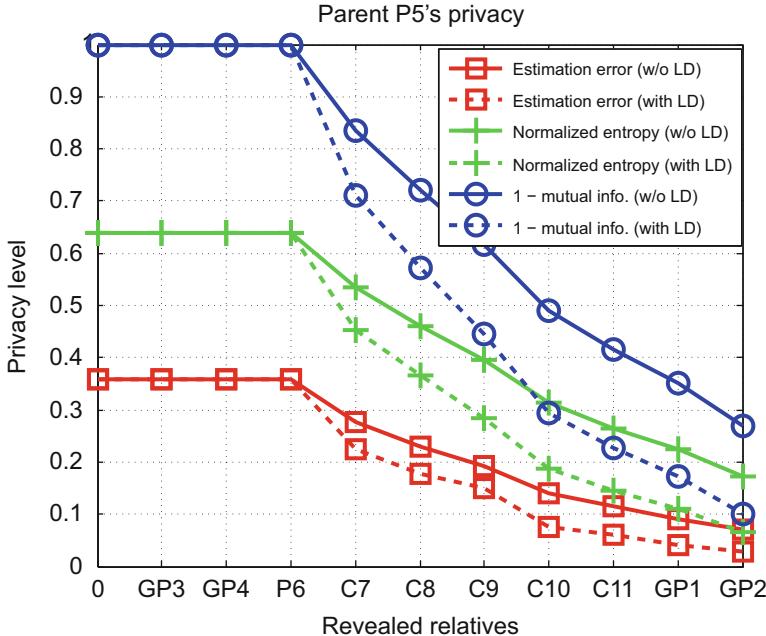


Fig. 18.3 Evolution of the genomic privacy of the parent (P5), with and without considering LD. For each family member, we reveal 50 randomly picked SNPs (among 100 SNPs in S), starting from the most distant family members, and the x -axis represents the exact sequence of this disclosure. Note that $x = 0$ represents the prior distribution, when no genomic data is revealed

and we keep revealing relatives until we reach his/her closest family members.⁴ We observe that individuals sometimes reveal different parts of their genomes (e.g., different sets of SNPs) on the Internet. Thus, we assume that for each family member (except for the target individual), the adversary observes 50 random SNPs from S only (instead of all the SNPs in S), and these sets of observed SNPs are different for each family member. In Fig. 18.3, we show the evolution of genomic privacy of one target individual (P5). We quantify the genomic privacy based on (i) attackers incorrectness (bottom plot), (ii) attacker’s uncertainty (middle plot), and (iii) an entropy-based metrics that quantifies the mutual dependence between the hidden genomic data that the adversary is trying to reconstruct (top plot). We observe that LD decreases genomic privacy, especially when few individuals’ genomes are revealed. As more family member’s genomes are observed, LD has less impact on the genomic privacy.

As we already mentioned, the Lacks family is just one (albeit famous) example. In the future (and already today), people of the same family might have very differ-

⁴The exact sequence of the family members (whose SNPs are revealed) is indicated for each evaluation.

ent opinions on whether to reveal genomic data, and this can lead to disagreement: relatives might have divergent perceptions of possible consequences. It is high time for the security research community to prepare itself for this formidable challenge. The genetics community is highly concerned about the fact that the proliferation of negative stories could potentially lead to a negative perception by the population and to tighter laws, thus hampering scientific progress in this field.

In order to prevent some of the aforementioned threats on the privacy of genomic data, we proposed several solutions to protect the privacy of such data in various domains. In the next section, we describe some of these solutions.

18.2 Solutions for Genomic Privacy

In this section, we summarize some of our efforts to protect the privacy of genomic data by focusing on privacy-preserving management of raw genomic data, privacy compliant use of genomic data in personalized medicine and research settings, resistance to brute-force attacks for storage of genomic data, and protecting kin genomic privacy.

18.2.1 Privacy-Preserving Management of Raw Genomic Data

Sequence alignment/map (SAM and its binary version BAM) files are the *de facto* standards used to store the aligned,⁵ raw genomic data generated by next-generation DNA sequencers and bioinformatic algorithms. There are hundreds of millions of short reads (each including between 100 and 400 nucleotides) in the SAM file of an individual. Typically, each nucleotide is present in several short reads in order to have sufficiently high coverage of each individual DNA.

In general, geneticists prefer storing aligned, raw genomic data of the patients (i.e., their SAM files), in addition to their variant calls (which include each nucleotide on the DNA sequence once, hence is much more compact) due to the following reasons: (i) Bioinformatic algorithms and sequencing platforms for variant calling are currently not yet mature, and hence geneticists prefer to observe each nucleotide in several short reads; (ii) If a patient carries a disease, which causes specific variations in the diseased cells (e.g., cancer), his or her DNA sequence in his/her healthy cells will be different from those diseased. Such variations can be misclassified as sequencing errors by only looking at the patient's variant calls (rather than his/her short reads). Furthermore, (iii) due to the rapid evolution of genomic research, geneticists do not know enough to decide which information

⁵Alignment is with respect to the reference genome, which is assembled by the scientists.

should really be kept and what is superfluous, hence they prefer to store all outcomes of the sequencing process as SAM files.

In Ayday et al. [4], we proposed a privacy-preserving system for the storage, retrieval, and processing of the SAM files. In a nutshell, the proposed scheme stores the encrypted SAM files of the patients at a *biobank* and it provides the requested range of nucleotides (on the DNA sequence) to a medical unit (for a genetic test) while protecting the patients' genomic privacy. It is important to note that the proposed scheme enables the privacy-preserving processing of the SAM files both for individual treatment (when the medical unit is embodied in a hospital) and for genetic research (when the medical unit is embodied in a pharmaceutical company).

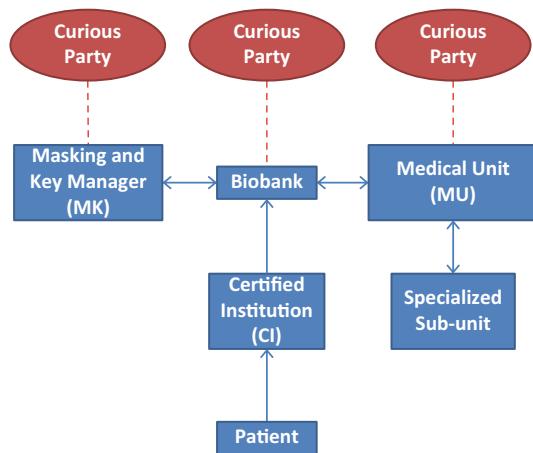
We assume that the sequencing and encryption of the genomes are done at a *certified institution* (CI), which is a trusted entity. We note that having such a trusted entity cannot be avoided as the sequencing has to be done at some institution to obtain the SAM files of the patients. Each part (position, cigar string, and content)⁶ of each short read (in the SAM file) is encrypted (via a different encryption scheme) after the sequencing, and encrypted SAM files of the patients are stored at a biobank. We assume that SAM files are stored at the biobank by using pseudonyms; this way, the biobank cannot associate the conducted genetic tests and the *medical unit* (MU), which conduct these tests, with the real identities of the patients. We note that a private company (e.g., cloud storage service) or the government could play the role of the biobank. There are potentially multiple MUs in the system, and each MU is an approved institution (by the medical authorities). Furthermore, we assume that an MU is a broad unit consisting of many sub-units (e.g., physicians or specialized clinics) that can potentially request nucleotides from any parts of a patient's genome.

The cryptographic keys of the patients are stored on a key manager by using the patient's pseudonym (which does not require the participation of the patient in the protocol). From here on, we assume the existence of a *masking and key manager* (MK) in the system to store the cryptographic keys of the patients. The MK can also be embodied in the government or a private company. The connection between these parties in the proposed protocol (along with the assumed threat model) is illustrated in detail in Fig. 18.4.

When the MU requests a specific range of nucleotides (on the DNA sequence of one or multiple patients), the biobank provides all the short reads that include at least one nucleotide from the requested range through the MK. During this process, the patient does not want to reveal his complete genome to the MU, to the biobank, or to the MK. Furthermore, it is not desirable for the biobank to learn the requested range of nucleotides (as the biobank can infer the nature of the genetic test from this requested range). Thus, we developed a privacy-preserving system for the retrieval of the short reads by the MU [4]. The proposed scheme provides the short reads that

⁶Position of a short read tells the position of the first nucleotide on the DNA sequence. Cigar string of a short read denotes the deletions and insertions on the short read. Content of a short read includes the nucleotides.

Fig. 18.4 Connections between the parties in the proposed protocol for privacy-preserving management of raw genomic data [4]



include the requested range of nucleotides to the MU without revealing the positions of these short reads to the biobank.

To achieve this goal, we first modify the structure of the genome by permuting the positions of the short reads, and then we use order preserving encryption (OPE) on the positions of the short reads (in the SAM file). OPE is a deterministic encryption scheme whose encryption function preserves numerical ordering of the plaintexts [1, 37]. Thus, OPE enables the encryption of the positions of the short reads and preserves the numerical ordering of the plaintext positions.

We prevent the leakage of extra information in the short reads to the MU by masking the encrypted short reads at the biobank (before sending them to the MU). As each short read includes between 100 and 400 nucleotides, some provided short reads might include information out of the MU's requested range of genomic data, as in Fig. 18.5. Similarly, some provided short reads might contain privacy-sensitive SNPs of the patient (which would reveal the patient's susceptibilities to privacy-sensitive diseases such as Alzheimer's), hence the patient might not give consent to reveal such parts, as in Fig. 18.6. To achieve this goal, we encrypt the content of the short reads by using stream cipher and mask certain parts of the encrypted short reads at the biobank, without decrypting them using an efficient algorithm. It is important to note that after the short reads are decrypted at the MU, the MU is not able to determine the nucleotides at the masked positions. This proposed system is very efficient and it has been adopted in real-life by bioinformatics companies.

18.2.2 Private Use of Genomic Data in Personalized Medicine

In Ayday et al. [6], we proposed a scheme to protect the privacy of users' genomic data while enabling medical units to access the genomic data in order to conduct medical tests or develop personalized medicine methods. In a medical test, a medical

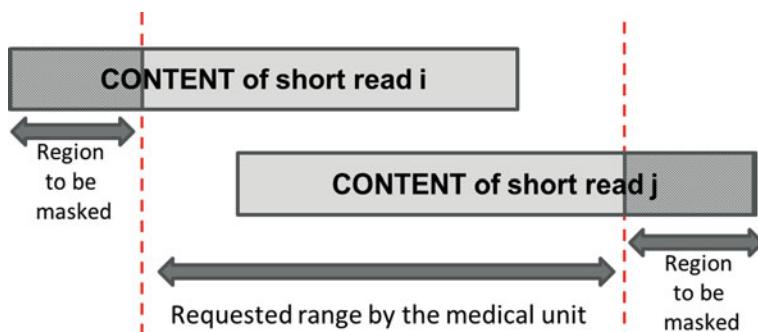
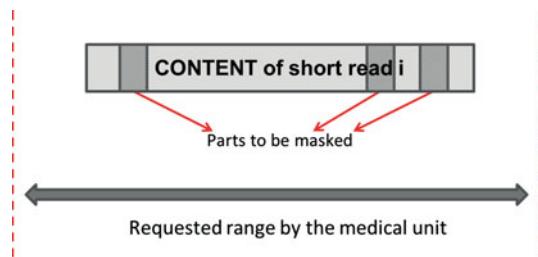


Fig. 18.5 Parts to be masked in the short reads for out-of-range content

Fig. 18.6 Parts to be masked in a short read based on patient's consent. The patient does not give consent to reveal the dark parts of the short read



unit checks for different health risks (e.g., disease susceptibilities) of a user by using specific parts of his genome. Similarly, to provide personalized medicine, a pharmaceutical company tests the compatibility of a user with a particular medicine. It is important to note that these genetic tests are currently done by different types of medical units, and the tools we propose in this work aim to protect the genomic privacy of the patients in such tests. In both medical tests and personalized medicine methods, in order to preserve his privacy, the user does not want to reveal his complete genome to the medical unit or to the pharmaceutical company. In addition, in some scenarios, it is the pharmaceutical companies who do not want to reveal the genetic properties of their drugs. To achieve these goals, we introduced the *privacy-preserving disease susceptibility test* (PDS) [6].

Most medical tests and personalized medicine methods (that use genomic data) involve a patient and a medical unit. In general, the medical unit can be a physician in a medical center (e.g., hospital), a pharmacist, a pharmaceutical company, or a medical council. In this study, we consider the existence of a curious entity in the medical unit as the potential attacker. That is, a medical unit might contain a disgruntled employee or it can be hacked by an intruder that is trying to obtain private genomic information about a patient (for which it is not authorized).

In addition, extreme precaution is needed for the storage of genomic data due to its sensitivity. Thus, we claim that a storage and processing unit (SPU) should be used to store the genomic data. We assume that the SPU is more “security-aware” than a medical unit, hence it can protect the stored genomic data against a hacker

better than a medical unit (yet, attacks against the SPU cannot be ruled out, as we discuss next). Recent medical data breaches from various medical units also support this assumption. Furthermore, instead of every medical unit individually storing the genomic data of the patients (in which case patients need to be sequenced by several medical units and their genomic data will be stored at several locations), a medical unit can retrieve the required genomic data belonging to a patient directly from the SPU. We note that a private company (e.g., cloud storage service), the government, or a non-profit organization could play the role of the SPU.

We assume that the SPU is an honest organization, but it might be curious. In other words, the SPU honestly follows the protocols and provides correct information to the other parties, however, a curious party at the SPU could access or infer the stored genomic data. Further, it is possible to identify a person only from his genomic data via phenotyping, which determines the observable physical or biochemical characteristics of an organism from its genetic makeup and environmental influences. Therefore, genomic data should be stored at the SPU in encrypted form. Similarly, apart from the possibility of containing a curious entity, the medical unit honestly follows the protocols. Thus, we assume that the medical unit does not make malicious requests from the SPU. We consider the following models for the attacker:

- A curious party at the SPU (or a hacker who breaks into the SPU), who tries to infer the genomic sequence of a patient from his stored genomic data. Such an attacker can infer the variants (i.e., nucleotides that vary between individuals) of the patient from his stored data.
- A semi-honest entity in the medical unit, who can be considered either as an attacker that hacks into the medical unit's system or a disgruntled employee who has access the medical unit's database. The goal of such an attacker is to obtain private genomic data of a patient for whom he or she is not authorized. The main resource of such an attacker is the results of the genetic tests that the patient undergoes.

For the simplicity of presentation, in the rest of this section, we will focus on a particular medical test (namely, computing genetic disease susceptibility). Similar techniques would apply for other medical tests and personalized medicine methods. In a typical genetic disease-susceptibility test, a *medical center* (MC) wants to check the susceptibility of a patient (P) for a particular disease X (i.e., the probability that patient P will develop disease X) by analyzing particular SNPs of the patient.⁷

For each patient, we propose to store only the *real SNPs* (around four million SNP positions on the DNA at which the patient has a mutation) at the SPU. At this point, it can be argued that these four million real SNPs (nucleotides) could be easily stored on the patient's computer or mobile device, instead of at the

⁷In this study, we only focused on the diseases which can be analyzed using the SNPs. We admit that there are also other diseases which depend on other forms of mutations or environmental factors.

SPU. However, we assert that this should be avoided due to the following issues. On one hand, types of variations in human population are not limited to SNPs, and there are other types of variations such as *copy-number variations* (CNVs), rearrangements, or translocations, consequently the required storage per patient is likely to be considerably more than only four million nucleotides. This high storage cost might still be affordable (via desktop computers or USB drives), however, genomic data of the patient should be available any time (e.g., for emergencies), thus it should be stored at a reliable source such as the SPU. On the other hand, leaving the patient's genomic data in his own hands and letting him store it on his computer or mobile device is risky, because his mobile device can be stolen or his computer can be hacked. It is true that the patient's cryptographic keys (or his authentication material) to access his genomic data at the SPU can also be stolen, however, in the case of a stolen cryptographic key, his genomic data (which is stored at the SPU) will still be safe. This can be considered like a stolen credit card issue. If the patient does not report that his keys are compromised, his genomic data can be accessed by the attacker.

It is important to note that protecting only the states (contents) of the patient's real SNPs is not sufficient in terms of his genomic privacy. As the real SNPs are stored at the SPU, a curious party at the SPU can infer the nucleotides corresponding to the real SNPs from their positions and from the correlation between the patient's potential SNPs and the real ones. That is, by knowing the positions of the patient's real SNPs, the curious party at the SPU will at least know that the patient has one or two minor alleles at these SNP positions (i.e., it will know that the corresponding SNP position includes either a real homozygous or heterozygous SNP), and it can make its inference stronger using the correlation between the SNPs.⁸ Therefore, in [6] we proposed to encrypt both the positions of the real SNPs and their states. We assume that the patient stores his cryptographic keys (public-secret key pair for asymmetric encryption, and symmetric keys between the patient and other parties) on his smart card (e.g., digital ID card). Alternatively, these keys can be stored at a cloud-based password manager and retrieved by the patient when required.

In short, the whole genome sequencing is done by a *certified institution* (CI) with the consent of the patient. Moreover, the real SNPs of the patient and their positions on the DNA sequence (or their unique IDs) are encrypted by the same CI (using the patient's public and symmetric key, respectively) and uploaded to the SPU, so that the SPU cannot access the real SNPs of the patient (or their positions). We are aware that the number of discovered SNPs increases with time. Thus, the patient's complete DNA sequence is also encrypted as a single vector file (via symmetric encryption using the patient's symmetric key) and stored at the SPU, thus when new SNPs are discovered, these can be included in the pool of the previously stored SNPs of the patient. We also assume the SPU not to have access to the real identities of the

⁸It is public knowledge that a real SNP includes at least one minor allele, and the curious party uses this background information in the attack.

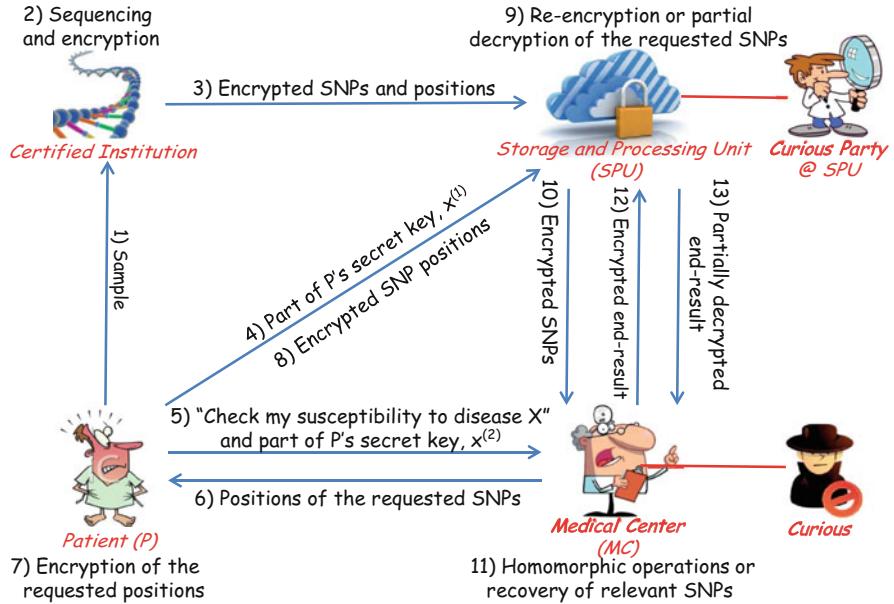


Fig. 18.7 Proposed privacy-preserving disease susceptibility test (PDS) [6]

patients and data to be stored at the SPU by using pseudonyms; this way, the SPU cannot associate the conducted genetic tests to the real identities of the patients.

Depending on the access rights of the MC, either (i) the MC computes $\text{Pr}(X)$, the probability that the patient will develop disease X by checking a subset of the patient's encrypted SNPs via homomorphic encryption techniques [7], or (ii) the SPU provides the relevant SNPs to the MC (e.g., for complex diseases that cannot be interpreted using homomorphic operations). These access rights are defined either jointly by the MC and the patient, or directly by the medical authorities. We note that homomorphic encryption lets the MC compute $\text{Pr}(X)$ using encrypted SNPs of patient P. In other words, the MC does not access P's SNPs to compute his disease susceptibility. We use a modification of the Paillier cryptosystem [2, 7] to support the homomorphic operations at the MC. We show our proposed protocol in Fig. 18.7.

Following the steps in the figure, initially, the patient (P) provides his sample (e.g., his blood or saliva) to the certified institution (CI) for sequencing. After sequencing, the CI first determines the positions of P's real SNPs and the set positions at which P has real SNPs. Then, CI encrypts the SNPs (with Paillier cryptosystem using the public key of the patients) and their positions (using the symmetric key shared between the patient and the CI). Next, the CI sends the encrypted SNPs and positions to the SPU and the patient provides a part of his secret key ($x^{(1)}$) to the SPU. This finalizes the initialization phase of the protocol. Then, the MC wants to conduct a susceptibility test on P for a particular disease X , and P provides the other part of his secret key ($x^{(2)}$) to the MC. The MC tells the patient the

positions of the SNPs that are required for the susceptibility test or requested directly as the relevant SNPs (but not the individual contributions of these SNPs to the test). The patient encrypts each requested position with the symmetric key and sends the SPU the encrypted positions of the requested SNPs. Next, the SPU re-encrypts the requested SNPs and sends them to the MC. MC computes P's total susceptibility for disease X by using the homomorphic properties (i.e., homomorphic addition and multiplication with a constant) of the modified Paillier cryptosystem. The MC sends the encrypted end-result to the SPU, which partially decrypts it using $x^{(1)}$ by following a proxy re-encryption protocol and sends it back to the MC. Finally, the MC decrypts the message received from the SPU by using $x^{(2)}$ and recovers the end-result.

Even though this proposed approach provides a secure algorithm, there is still a privacy risk in case the MC tries to infer the patient's SNPs from the end-result of a test. In [6], we also showed that such an attack is indeed possible and one way to prevent such an attack is to obfuscate the end-result before providing it to the MC. Obviously, this causes a conflict between privacy and utility and this conflict is still a hot research topic for genomic privacy.

In a follow up work [5], we also proposed a system for protecting the privacy of individuals' sensitive genomic, clinical, and environmental information, while enabling medical units to process it in a privacy-preserving fashion in order to perform disease risk tests. We introduced a framework in which individuals' medical data (genomic, clinical, and environmental) is stored at a storage and processing unit (SPU) and a medical unit conducts the disease risk test on the encrypted medical data by using homomorphic encryption and privacy-preserving integer comparison. The proposed system preserves the privacy of the individuals' genomic, clinical, and environmental data from a curious party at the SPU and from a curious party (e.g., a hacker) at the medical unit when computing the disease risk. We also implemented the proposed system and showed its practicality via a complexity evaluation.

The general architecture of the proposed system is illustrated in Fig. 18.8. In summary, the patient provides his sample for sequencing to the CI. Meanwhile, he also provides his clinical and environmental data to the SPU and the MU.⁹ The CI is responsible for sequencing and encryption of the patient's genomic data. Then, the CI sends the encrypted genomic data to the SPU. Finally, the privacy-preserving computation of the disease risk takes place between the MU and the SPU.

18.2.3 Private Use of Genomic Data in Research

The past years have witnessed substantial advances in understanding the genetic bases of many common phenotypes of biomedical importance. Such an evolution in

⁹Depending on the privacy-sensitivity of the clinical and environmental data, the patient can choose which clinical and environmental attributes to reveal to the MU, and which ones to encrypt and keep at the SPU.

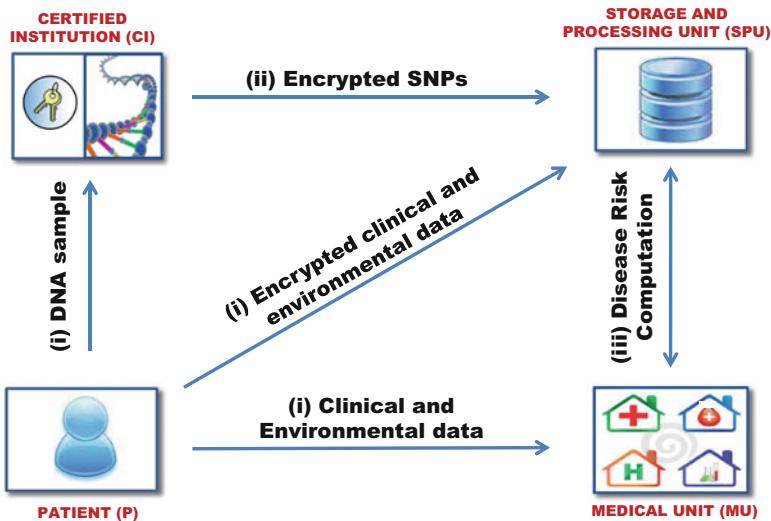


Fig. 18.8 Proposed system model for the privacy-preserving computation of the disease risk [5]

the medical field has pushed companies like Google to set up new infrastructures (e.g., Google Genomics [17]) to store, process and share genetic data at a large scale. Genome-wide association studies (GWAS) have become a popular method to investigate the relationship between the genomic variation and several diseases. They represent a starting point on the journey for translating this knowledge into clinics and they pave the way for personalized medicine, which is expected to have an unprecedented impact for clinical care by enabling treatment of diseases based on the genomic makeups of the individuals.

Even though much emphasis is given to GWAS, replication studies and fine-mapping of associated regions (which are both based on the *a priori* knowledge generated with GWAS) are crucial to identify true positive associations and variants that are causal for a phenotype. Replication studies are investigations performed in independent cohorts to validate variants identified by GWAS. Fine-mapping studies are useful in the post-GWAS phase when a few associations have been convincingly demonstrated and exhaustive work has to be performed to identify the actual causative variants. Additionally, it is becoming much more frequent to investigate multiple phenotypes across the same set of patient data, in so called genome-wide association studies (PheWAS), which allow researchers to better understand the genetic architecture of complex traits and gain insights into disease mechanisms.

As genetic association studies depend on a large amount of genomic-phenomic data, strong privacy guarantees are required in order to protect the sensitive health information of individuals and, thus, facilitate the pace of genomic research by encouraging people to participate in such studies knowing that their privacy is protected. As discussed, genomic data includes privacy-sensitive information about

an individual, such as his ethnicity, kinship, and predisposition to specific diseases. Leakage of such information may cause genetic discrimination or blackmail. Similarly, phenotype data of individuals is also sensitive as it includes an individual's disease status and identifiers. Even though standard anonymization techniques can be used to publish phenotype data (with decreased accuracy), they have proved to be ineffective for genomic data [18, 21]. Hence, more sophisticated privacy-enhancing technologies have to be developed.

In Raisaro et al. [38], we proposed a privacy-preserving technique to conduct replication and fine-mapping genetic association studies.¹⁰ We note that our solution is flexible enough to be generalized and to ensure privacy protection in different applications of the medical research field. Increasingly, large-scale data sets are being generated and applied in the medical setting, including proteomic, transcriptomic and metabolomic data. By recombining the building blocks of our privacy-preserving algorithm, the proposed architecture can easily support also secure analyses of multiple 'omics datasets for personalized medicine methods as proposed in Ayday et al. [6].

Existing techniques to conduct association studies in a privacy-preserving way include (i) adding noise to the result of the study to satisfy differential privacy [26, 44] (e.g., when the study is done at a trusted database and only the results of the study is shared with the researchers), and (ii) cryptographic techniques, such as using homomorphic encryption [28, 29] (e.g., when genomic data is shared with the researchers and the study is done by the researchers). Techniques in the former category reduce the utility of genomic data, and hence are criticized by genomic researchers, while cryptographic solutions enable computing exact answers with some computational and storage overhead [11, 14]. Our proposed technique falls into the latter category. However, as opposed to the existing crypto-based works, our proposed method in [38] (i) stores each participant's genotype and phenotype data encrypted by his own cryptographic key, (ii) addresses, for the first time in a privacy-preserving way, the problem of population stratification, and (iii) is highly parallelizable. We emphasize that our method, by storing each participant's data encrypted by his own key, avoids a single point of failure in the system. If a key is leaked or hacked, only the data of a single participant is compromised and other participants' data is still protected. Conversely, previous solutions assume that all participants' data is stored encrypted under the same key, therefore, they are less secure as the leakage of such a key could jeopardize the entire system.

In a nutshell, we developed an efficient privacy-preserving algorithm for genetic association studies on encrypted genotypes and phenotypes stored in a centralized dataset. Our solution addresses the pervasive challenge of dataset stratification by inferring, in a privacy-preserving way, the ancestry of each subject in the dataset. Identification of dataset stratification represents a crucial preprocessing step of genetic association studies to avoid spurious associations due to systematic ancestry

¹⁰Our solution may also be used for GWAS, but it better scales for replication/fine-mapping association studies which are based on the *a priori* knowledge generated with GWAS.

differences within and between sample populations. Furthermore, our algorithm automatically generates case and control groups (i.e., two sets of individuals differing in one or more phenotypic traits) and outputs only the final result of the association study without leaking any information of the intermediate steps of the computation. We prove the security of the proposed technique and assess its performance with an implementation on real data. We also propose a MapReduce implementation as a proof-of-concept of parallelization.

One real-life application of the proposed technique is clinical studies conducted by pharmaceutical companies in collaboration with national biobanks. The goal of these studies is to assess the effectiveness of a treatment (or effect of a drug) for a certain group of people. In such a scenario, we can assume the biobank stores the encrypted genotypes and phenotypes of a set of individuals. Then, a pharmaceutical company can run a privacy-preserving genetic association study to identify *in a few hours* the set of genetic variants that influence the efficacy of the treatment. Today, these types of pharmacogenetic studies are performed through methods that are not privacy-preserving. Since biobanks cannot release data without the explicit consent from the participants or a special approval from an ethics committee, a pharmacogenetic study can require months to be completed. Therefore, the proposed technique not only preserves the privacy of the individuals' sensitive health-related data, but it also accelerates the pace of genomic research.

In general, genetic association studies involve a cohort of participants (P), who provide upon consent their genotype and phenotype information for research purposes, and a medical unit (MU) that performs the association study on this cohort. As discussed, the MU can be either a pharmaceutical company willing to conduct a clinical trial for a particular drug, or a research institution willing to test the association between some single nucleotide variations (SNVs) of significant interest and complex phenotypic traits. As shown in Fig. 18.9, the proposed system in [38] includes a certified institution (CI) and a centralized storage and processing unit (SPU), along with the P and the MU. The CI is responsible for (i) recruiting the participants for association studies, (ii) genotyping their genome (i.e., identifying and extracting their genetic variations), (iii) collecting their phenotype information, (iv) encrypting the data, and (v) generating and distributing the cryptographic keys between the parties.

We assume, for efficiency and security, the storage of encrypted genotypes and phenotypes to be at the SPU. That is, instead of several MUs storing the same large amount of genomic and phenomic data, the information of each participant is stored at a centralized SPU and, upon request, made accessible (for association studies) to different MUs. Storing genotype and phenotype information at the SPU also enables (i) data from multiple hosts to be pooled into a single and centralized repository, and (ii) genomic association studies to be conducted on an amount of data often beyond the capability of a sole researcher or institution. The purpose of such an architecture is to overcome the main limiting factor of association studies, i.e., insufficient sample size, as the individual effect of genomic differences is usually small, and large sample sizes are required in order to increase the sensitivity of statistical tests and data-mining techniques. As before, a private company (e.g.,

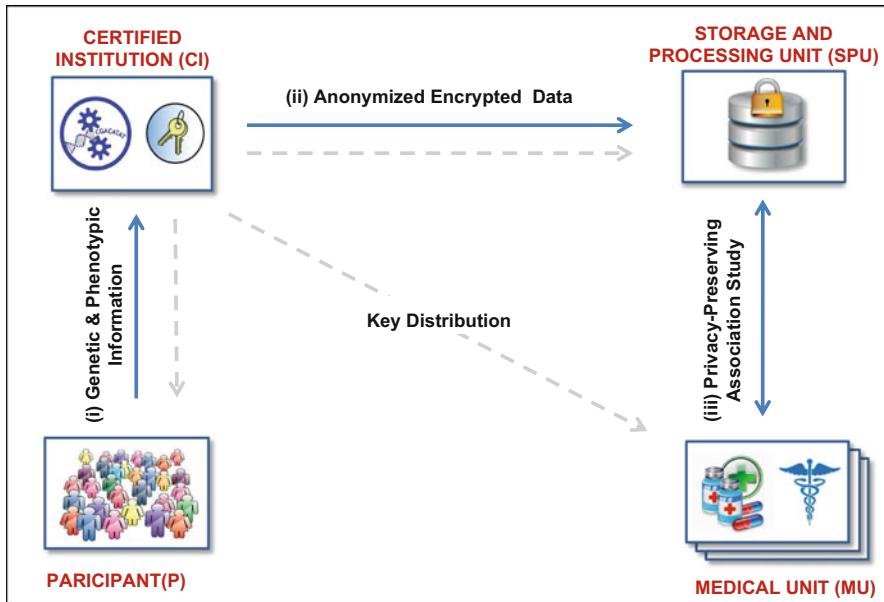


Fig. 18.9 System model for private use of genomic data in research setting [38]: participants (P), certified institution (CI), storage and processing unit (SPU), and medial units (MU)

cloud storage service), the government, or a non-profit organization can play the role of the SPU. The proposed algorithm for privacy-preserving genetic association studies takes place between the MU and the SPU.

The proposed solution in [38] can be summarized as follows. First, the participants provide to the CI their biological sample for genotyping, along with their phenotype information. Then, the CI encrypts each participant's information and sends it to the SPU. Finally, after a preprocessing phase for ancestry inference, the privacy-preserving genetic association study takes place between the MU and the SPU through a secure two-party protocol (using the homomorphic properties of the Paillier cryptosystem and some SMC protocols between the MU and the SPU). In such a protocol, the MU specifies the input parameters to the SPU and obtains only the allele frequencies for the two study groups.

18.2.4 Coping with Weak Passwords for the Protection of Genomic Data

Appropriately designed cryptographic schemes can preserve the data utility, but they provide security based on assumptions about the computational limitations of adversaries. Hence, they are vulnerable to brute-force attacks when these

assumptions are incorrect or erode over time. Given the longevity of genomic data, serious consequences can result. Compared with other types of data, genomic data has especially long-term sensitivity. A genome is (almost) stable over time and thus needs protection over the lifetime of an individual and even beyond, as genomic data is correlated between the members of a single family. It has been shown that the genome of an individual can be probabilistically inferred from the genomes of his or her family members [23].

In many situations, though, particularly those involving direct use of data by consumers, keys are weak and vulnerable to brute-force cracking *even today*. Users' tendency to choose weak passwords is widespread and well documented [12]. This problem arises in systems that employ password-based encryption (PBE), a common approach to protection of user-owned data.

Recently, Juels and Ristenpart introduced a new theoretical framework for encryption called *honey encryption* (HE) [27]. Honey encryption has the property that when a ciphertext is decrypted with an incorrect key (as guessed by an adversary), the result is a plausible-looking yet incorrect plaintext. Therefore, HE gives encrypted data an additional layer of protection by serving up fake data in response to every incorrect guess of a cryptographic key or password. Notably, HE provides a hedge against brute-force decryption in the long term, giving it a special value in the genomic setting.

However, HE relies on a highly accurate distribution-transforming encoder (DTE) over the message space. Unfortunately, this requirement jeopardizes the practicality of HE. To use HE in any scenario, we have to understand the corresponding message space quantitatively, that is, the precise probability of every possible message. When messages are not uniformly distributed, characterizing and quantifying the distribution is a highly non-trivial task. Building an efficient and precise DTE is the main challenge when extending HE to a real use case.

In Huang et al. [22], we proposed to address the problem of protecting genomic data by combining the idea of honey encryption with the special characteristics of genomic data in order to develop a secure genomic data storage (and retrieval) technique that is (i) robust against potential data breaches, (ii) robust against a computationally unbounded adversary, and (iii) efficient.

In the original HE paper [27], Juels and Ristenpart propose specific HE constructions that rely on existing generation algorithms (e.g., for RSA private keys), or operate over very simple message distributions (e.g., credit card numbers). These constructions, however, are inapplicable to plaintexts with considerably more complicated structure, such as genomic data. Thus, substantially new techniques are needed in order to apply HE to genomic data. Additional complications arise when the correlation between the genetic variants (on the genome) and phenotypic side information are taken into account. Our work in [38] is devoted mainly to addressing these challenges.

We proposed a scheme called GenoGuard. In GenoGuard [38], genomic data is encoded to generate a *seed* value, the seed is encrypted under a patient's password,¹¹ and stored at a centralized biobank. We propose a novel tree-based technique to efficiently encode (and decode) the genomic sequence in order to meet the special requirements of honey encryption. Legitimate users of the system can retrieve the stored genomic data by typing their passwords.

A computationally unbounded adversary can break into the biobank protected by GenoGuard, or remotely try to retrieve the genome of a victim. The adversary could exhaustively try all the potential passwords in the password space for any genome in the biobank. However, for each password he tries (thanks to our encoding phase), the adversary will obtain a plausible-looking genome without knowing whether it is the correct one. We also consider the case when the adversary has side information about a victim (or victims) in terms of his physical traits. In this case, the adversary could use genotype-phenotype associations to determine the real genome of the victim. GenoGuard is designed to prevent such attacks, hence it provides protections beyond the normal guarantees of HE.

We show the main steps of the GenoGuard protocol in Fig. 18.10. We represent the patient and the user as two separate entities, but they can be the same individual, depending on the application.

GenoGuard is highly efficient and can be used by the service providers that offer DTC services (e.g., 23andMe) to securely store the genomes of their customers. It can also be used by medical units (e.g., hospitals) to securely store the genomes of

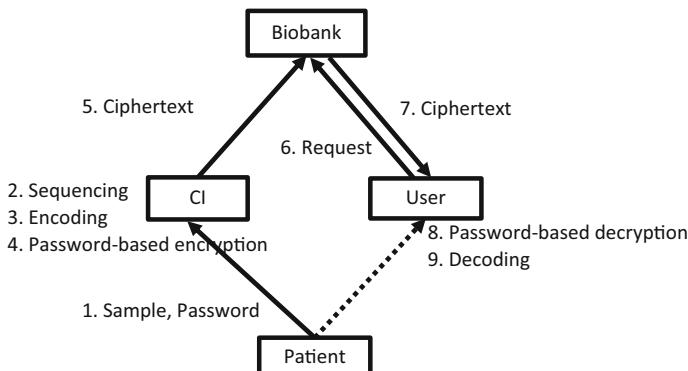


Fig. 18.10 GenoGuard protocol [38]. A patient provides his biological sample to the CI, and chooses a password for honey encryption. The CI does the sequencing, encoding and password-based encryption, and then sends the ciphertext to the biobank. During a retrieval, a user (e.g., the patient or his doctor) requests for the ciphertext, decrypts it and finally decodes it to get the original sequence

¹¹A patient can choose a low-entropy password that is easier for him/her to remember, which is a common case in the real world [12].

patients and to retrieve them later for clinical use. The general protocol in Fig. 18.10 can work in a healthcare scenario without any major changes. In this scenario, a patient wants a medical unit (e.g., his doctor) to access his genome and perform medical tests. The medical unit can request for the encrypted seed on behalf of (and with consent from) the patient. Hence, there is a negotiation phase that provides the password to the medical unit. Such a phase can be completed automatically via the patient's smart card (or smart phone), or the patient can type his password himself. In this setup, the biobank can be a public centralized database that is semi-trusted. Such a centralized database would be convenient for the storage and retrieval of the genomes by several medical units.

For direct-to-customer (DTC) services, the protocol needs some adjustments. For instance, Counsyl¹² and 23andme¹³ provide their customers various DTC genetic tests. In such scenarios, the biobank is the private database of these service providers. Thus, such service providers have the obligation to protect customers' genomic data in case of a data breach. In order to perform various genetic tests, the service providers should be granted permission to decrypt the sequences on their side, which is a reasonable relaxation of the threat model because customers share their sequences with the service providers. Therefore, steps 8 and 9 in Fig. 18.10 should be moved to the biobank. A user who requests a genetic test result logs into the biobank system, provides the password for password-based decryption and asks for a genetic test on his sequence. The plaintext sequence is deleted after the test.

18.2.5 Protecting Kin Genomic Privacy

In Humbert et al. [24], we presented a genomic-privacy preserving mechanism (GPPM) for reconciling people's willingness to share their genomes (e.g., to help research¹⁴) with privacy. Our GPPM acts at the individual data level, not at the aggregate data (or statistical) level like in [26]. Focusing on the most relevant type of variants (the SNPs), we study the trade-off between the usefulness of disclosed SNPs (utility) and genomic privacy. We consider an individual who wants to share his genome, yet who is concerned about the subsequent privacy risks for himself and his family. Thus, we design a system that maximizes the disclosure utility but does not exceed a certain level of privacy loss within a family, considering (i) kin genomic privacy, (ii) personal privacy preferences (of the family members), (iii) privacy sensitivities of the SNPs, (iv) correlations between SNPs, and (v) the research utility of the SNPs. The proposed GPPM in [24] can automatically evaluate the privacy risks of all the family members and decide which SNPs to disclose. To achieve

¹²<https://www.counsyl.com/>.

¹³<https://www.23andme.com/>.

¹⁴<http://opensnp.wordpress.com/2011/11/17/first-results-of-the-survey-on-sharing-genetic-information/>.

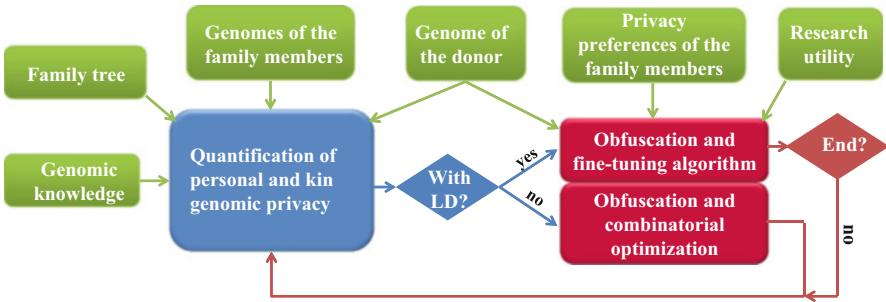


Fig. 18.11 General protection framework. The GPPM [24] takes as inputs (i) the privacy levels of all family members, (ii) the genome of the donor, (iii) the privacy preferences of the family members, and (iv) the research utility. First, correlations between the SNPs (LD) is not considered in order to use combinatorial optimization. Note that we go only once through this box. Then, LD is used and a fine-tuning algorithm is used to cope with non-linear constraints. The algorithm outputs the set of SNPs that the donor can disclose

this goal, it relies on probabilistic graphical models and combinatorial optimization. Our results indicate that, given the current data model, genomic privacy of an entire family can be protected while an appropriate subset of genomic data can be made available.

In order to mitigate attribute-inference attacks and protect genomic and health privacy, the GPPM relies upon an *obfuscation mechanism*. In practice, obfuscation can be implemented by adding noise to the SNP values, by injecting fake SNP values, by reducing precision, or by simply hiding the SNP values. In this work, we choose SNP hiding, essentially because the genomic research community would not receive other options positively. Indeed, genetic researchers are very reluctant about adding noise or fake data, notably because of the huge investment they make to increase (sequencing) accuracy. We assume one family member, at a given time, who wants to disclose his SNPs and to guarantee a minimum privacy level for him and his family. Figure 18.11 provides an overview of the proposed GPPM in [24].

For clarity of presentation, we focus on one family whose members are defined by the set \mathbf{F} ($|\mathbf{F}| = n$). We assume that there is only one donor D who makes the decision to share his genome at a given time. His relatives might have already publicly shared some of their genomic data on the Internet. D takes this into account when he makes his own disclosure decision. We let \mathbf{S} ($|\mathbf{S}| = m$) be the set of SNP IDs. Its cardinality m can go up to 50 million, as this is currently the approximate number of SNPs in the human population. In practice, however, people put online (e.g., on OpenSNP) up to one million of the most significant SNPs. We let $\mathbf{X}^D = \{x_j^D : j \in \mathbf{S}\}$ represent the set of SNPs of D (x_j^D is the value of SNP j of the donor D), that are all initially undisclosed. Finally, we let $\mathbf{y}^D = \{y_j^D : j \in \mathbf{S}\}$ represent the decision vector of D , where $y_j^D = 1$ means the corresponding SNP will be disclosed, and $y_j^D = 0$ means x_j^D will remain hidden.

We express the privacy constraints of a family member both in terms of genomic and health privacy. Our framework can account for different privacy preferences for different family members, SNPs, and diseases. For all $i \in \mathbf{F}$, $j \in \mathbf{S}$, we define the privacy sensitivity of a SNP j for individual i as s_j^i . We can set the s_j^i 's to be equal by default. Then, an individual willing to personalize his privacy preferences may further define his own privacy sensitivities regarding specific SNPs based on his privacy concerns regarding, e.g., certain phenotypes. The most well-known example of such a scenario is the case of James Watson, co-discoverer of DNA, who made his whole DNA sequence publicly available, with the exception of one gene known as Apolipoprotein E (ApoE), one of the strongest predictors for the development of Alzheimer's disease.¹⁵ We let the sets \mathbf{P}_s^i and \mathbf{P}_d^i include the privacy-sensitive SNP IDs and privacy-sensitive diseases of individual i , respectively. We represent the tolerance to the genomic-privacy loss of individual i as $\text{Pri}(i, \mathbf{P}_s^i)$, and the tolerance to the health-privacy loss of individual i regarding disease $d \in \mathbf{P}_d^i$ as $\text{Pri}(i, d)$. These tolerance values represent the maximum privacy loss (after the disclosure of D 's SNPs) that an individual would bear. By considering the privacy losses instead of the absolute privacy levels, we ensure that the donor will more likely reveal a SNP whose value is already well inferred by the attacker before donor's disclosure (e.g., by using SNPs previously shared by the donor's relatives). Note that these tolerance values can always be updated for any new family member willing to disclose his genome. Finally, the utility function is a non-decreasing function of the norm of \mathbf{y}^D , as the knowledge of more SNPs can only help genomic research. We define u_j to be the utility provided by SNP j . Note that, in practice, the utility of the SNPs can be determined by the research authorities and can vary based on the study.

The donor faces an optimization problem: How to maximize research utility while protecting his own and his relatives' genomic and health privacy. First, the objective function is formally defined as $\sum_{j \in \mathbf{S}} u_j y_j^D$. Then, privacy constraints are defined, for each individual, as the sum of privacy losses induced by the donor's disclosure over all SNPs. This sum must be capped by the respective privacy loss tolerances of all family members. Formally, for all individuals $i \in \mathbf{F}$ and SNPs $j \in \mathbf{S}$, the privacy loss induced by the disclosure of x_j^D is defined as $(E_j^i(y_j^D = 0) - E_j^i(y_j^D = 1))$. Note here that the privacy loss at a given SNP j for any relative is only affected by the donor's decision y_j^D regarding SNP j but no other SNP $k \neq j$, meaning that LD correlations are not taken into account. Finally, note that if an individual i has already revealed his SNP j , i.e., $x_j^i \in \mathbf{X_O}$, the privacy loss at this SNP for i is zero, because $E_j^i(y_j^D = 0) = E_j^i(y_j^D = 1) = 0$. For all $i \in \mathbf{F}$, $j \in \mathbf{S}$, the privacy weight p_j^i is defined as

$$p_j^i = s_j^i \times (E_j^i(y_j^D = 0) - E_j^i(y_j^D = 1)). \quad (18.1)$$

¹⁵Later researchers have used correlations in the genome to unveil Watson's predisposition to Alzheimer's [35]. In this work, we also consider such correlations.

Clearly, p_j^i at a given SNP j can be different for each family member, depending on how close he is from the donor in the family tree, on the actual values x_j^i and x_j^D of his and the donor's SNPs, and on his sensitivity. Note that $s_j^i = 0 \forall j \notin \mathbf{P}_s^i$.

We can now define the linear optimization problem as

$$\begin{aligned} & \underset{\mathbf{y}^D}{\text{maximize}} \quad \sum_{j \in \mathbf{S}} u_j y_j^D \\ & \text{subject to} \quad \sum_{j \in \mathbf{P}_s^i} p_j^i y_j^D \leq \text{Pri}(i, \mathbf{P}_s^i), \forall i \in \mathbf{F} \\ & \quad \sum_{k \in \mathbf{S}_d} p_k^i y_k^D \leq \text{Pri}(i, d), \forall d \in \mathbf{P}_d^i, \forall i \in \mathbf{F} \\ & \quad y_j^D \in \{0, 1\}, \forall j \in \mathbf{S}, \end{aligned} \tag{18.2}$$

where \mathbf{S}_d is the set of SNPs that are associated with disease d .

Our optimization problem is very similar to the multidimensional knapsack problem [15]. We decide to follow the branch-and-bound method proposed by Shih [40], because it finds the optimal solution, represents a good trade-off between time and storage space, and allows for the extension of the algorithm to null and negative (privacy) weights. However, the LD correlations between the SNPs are not considered in the above optimization problem in order for the constraints to remain linear. Therefore, after getting the initial results from the linear optimization problem, we use a fine-tuning algorithm in order to decide to reveal less or more SNPs when LD is also considered.

18.3 Future Research Directions

Advances in genomics will soon result in large numbers of individuals having their genomes sequenced and obtaining digitized versions thereof. This poses a wide range of technical problems, which we explore below [3].

Storage and Accessibility: Genome at Rest Due to its sensitivity and size (about 3.2 billion nucleotides), one key challenge is where and how a digitized genome should be stored. It is reasonable to assume that an individual who requests (and likely pays for) genome sequencing should own the result, as is already the case with any other personal medical results and information. This raises numerous issues, including:

- Should the genome be stored on one's personal devices, e.g., a PC or a smartphone? If so, what, if any, special hardware security features (e.g., tamper-resistance) are needed?
- Can it be outsourced to a cloud provider?

- Should the sequencing facility keep an escrowed copy of the genome?
- Should it be entrusted to one's personal physician and/or health insurance provider?
- How is it to be stored: in the clear or encrypted? If the latter, where are encryption keys generated: at the lab? at owner's premises? at the cloud provider? Where are these keys stored?
- How to guarantee integrity and authenticity of the digitized genome?
- Should backups be made? If so, how often and where can copies be kept?
- How can one erase a genome securely?
- Should an individual periodically re-sequence their genome to take advantage of more accurate technology?

Privacy: Genome in Action Given the genome's sensitivity, an individual should, ideally, never disclose any information contained therein. However, this would prevent the access to any genomic application that cannot be entirely and securely performed *in situ*, i.e., within a secure perimeter of one's own personal device. In principle, this might be possible if operations are performed in some standardized and certified form. For example, if testing for a genetic disease requires matching a well-known pattern in some approximate location in the genome, that pattern and its parameters can be certified by some trusted agency (such as the US Food and Drug Administration). Thus, an individual could be assured that a legitimate test for a specific genetic disease is being conducted and the result is clearly communicated to that individual; the latter would then have the option to keep the result private.

At the same time, it is hard to foresee the range and complexity of future genetic operations: some (future) tests might be too computationally complex to be performed within the confines of a personal device. Furthermore, some genetic testing would probably involve multiple genomes, e.g., when tracing origins of some conditions, siblings or parents/children might need to be tested together. Similarly, in assessing risks of genetic conditions for future progeny, both prospective parents have to be tested. Also, some genetic tests constitute intellectual property of a pharmaceutical/biomedical company (which needs to be protected) [8, 19, 34].

As soon as genomic information leaves the (virtual) hands of its owner, purely technical approaches to privacy become insufficient. Legal and professional guidelines are certainly needed to govern how information is transmitted, stored, processed, and eventually disposed of on the receiving end, e.g., by the physician, hospital, pharmacist or medical lab.

Long-term Data Protection Even if genomes are encrypted, encryption schemes considered strong today might gradually weaken in the long term, whereas genome sensitivity does not dissipate over time. It is not too far-fetched to imagine that a third-party in possession of an encrypted genome might be able to decrypt it years or decades later. For instance, the Advanced Encryption Standard (AES) scheme supports key lengths up to 256 bits—a key length estimated by NIST, following Moore's law, to be secure several years after 2030. However, computational breakthroughs or unforeseen weaknesses might allow breaking the encryption earlier than

expected. Also, even leakage of a long-deceased individual's genome could affect genomic privacy of that person's living progeny.

Assuming that it cannot be copied, an encrypted genome could be periodically re-encrypted. Alternatively, one could split the genome (e.g., by using secret-sharing techniques [39]), and partition it among several providers. However, this opens the problem of efficient reassembly of the genome for various operations as well as how to guarantee non-collusion between providers.

Accuracy and Accountability Computational genomic tests should guarantee accuracy at least equivalent to that of their current analog *in vitro* counterparts. For example, a software implementation of the paternity test should offer at least the same confidence as its *in vitro* counterpart currently admissible in a court of law. Also, computational tests should aim at accountability, e.g., by providing lasting guarantees of correctness for both execution and input information.

Efficiency Computational genomic tests should incur minimal communication and computational costs. Minimality in this setting is relative to the context of such tests. For instance, patients may be inclined (and accustomed) to wait several days to obtain results of genetic tests that concern their health. However, in the computational setting, long running times on personal devices might hinder the real-world practicality of these tests (besides negating one of the main motivations for computational tests).

Usability Computational genomic tests that involve end-users should be usable by, and meaningful to, regular non-tech-savvy individuals. This translates into non-trivial questions, such as: how much understanding should be expected from a user running a test? What information (and at what level of granularity) should be presented to the user as part of a test and as its outcome? Do privacy perceptions and concerns experienced by patients match those expected by the scientific community? Some users might be willing to forego their genomic privacy in certain cases. For instance, one may think that patients will be likely to reveal their genomes to their medical doctors (and hence trade off privacy of their genomes) to enable tests that can save them from, e.g., cancer. In contrast, in the case of online services or pharmaceuticals, an individual might not wish to forgo privacy. However, very few efforts (e.g., [13]) have focused on users' concerns, thus prompting the need for ethnographic studies. Also, there remains an open problem of how to effectively communicate to the users potential privacy risks associated with genomic information and its disclosure.

Large-scale Research on Human Genomes Potential privacy, legal, and ethical concerns appear to conflict with large-scale research on human genomes, such as Genome-Wide Association Studies (GWAS). However, large scale studies are needed to discover associations between genetic make-up and medical conditions. One current trend is to store donors' genomes in the cloud and use analytics techniques running on powerful computer clusters. Once again, this prompts many privacy and legal concerns.

18.4 Conclusion

In this chapter, focusing on the work carried out by EPFL/LCA1 and Bilkent University, we first discussed some threats for genomic privacy. Then, we described some of our solutions to protect genomic privacy. Namely, we focused on privacy-preserving management of raw genomic data, privacy compliant use of genomic data in personalized medicine and research settings, resistance to brute-force attacks and protecting kin genomic privacy. Finally, we discussed some future research directions. More information on this topic can be found at: <https://genomeprivacy.org>

Acknowledgements The authors would like to express their gratitude to Mathias Humbert, Jean Louis Raisaro, Zhicong Huang, Emiliano De Cristofaro, Gene Tsudik, Jacques Fellay, Amalio Telenti and Paul McLaren.

References

1. Agrawal, R., Kiernan, J., Srikant, R., Xu, Y.: Order preserving encryption for numeric data. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, pp. 563–574 (2004)
2. Ateniese, G., Fu, K., Green, M., Hohenberger, S.: Improved proxy re-encryption schemes with applications to secure distributed storage. ACM Trans. Inf. Syst. Secur. **9**, 1–30 (2006)
3. Ayday, E., Cristofaro, E.D., Tsudik, G., Hubaux, J.-P.: Whole genome sequencing: revolutionary medicine or privacy nightmare. IEEE Computat **48**(2), pp. 58–66 (2015)
4. Ayday, E., Raisaro, J.L., Hengartner, U., Molyneaux, A., Hubaux, J.-P.: Privacy-preserving processing of raw genomic data. In: Proceeding of 8th International Workshop on Data Privacy Management (DPM). Egham, UK (2013)
5. Ayday, E., Raisaro, J.L., McLaren, P.J., Fellay, J., Hubaux, J.-P.: Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data. In: Proceedings of USENIX Security Workshop on Health Information Technologies (HealthTech) (2013)
6. Ayday, E., Raisaro, J.L., Rougemont, J., Hubaux, J.-P.: Protecting and evaluating genomic privacy in medical tests and personalized medicine. In: CM Workshop on Privacy in the Electronic Society (WPES). Berlin, Germany (2013)
7. Bresson, E., Catalano, D., Pointcheval, D.: A simple public-key cryptosystem with a double trapdoor decryption mechanism and its applications. In: Proceedings of Asiacrypt (2003)
8. Caulfield, T., Cook-Deegan, R.M., Kieff, F.S., Walsh, J.P.: Evidence and anecdotes: an analysis of human gene patenting controversies. Nat. Biotechnol. **24**(9), pp. 1091–1094 (2006)
9. Clayton, D.: On inferring presence of an individual in a mixture: a bayesian approach. Biostatistics **11**(4), 661–673 (2010)
10. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al.: Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science **327**(5961), 78–81 (2010)
11. Erlich, Y., Narayanan, A.: Routes for breaching and protecting genetic privacy. Nat. Rev. Genet. **15**(6), 409–421 (2014)
12. Florencio, D., Herley, C.: A large-scale study of web password habits. In: Proceedings of the 16th International Conference on World Wide Web, WWW '07, pp. 657–666. ACM, New York (2007). doi:[10.1145/1242572.1242661](https://doi.acm.org/10.1145/1242572.1242661). url:<http://doi.acm.org/10.1145/1242572.1242661>

13. Francke, U., Dijamco, C., Kiefer, A.K., Eriksson, N., Moiseff, B., Tung, J.Y., Mountain, J.L.: Dealing with the unexpected: consumer responses to direct-access BRCA mutation testing. *PeerJ* **1** (2013)
14. Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., Ristenpart, T.: Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. In: Proceedings of the 23rd USENIX Security Symposium (2014)
15. Fréville, A.: The multidimensional 0–1 knapsack problem: an overview. *Eur. J. Oper. Res.* **155**(1), 1–21 (2004)
16. Gitschier, J.: Inferential genotyping of γ chromosomes in latter-day saints founders and comparison to Utah samples in the hapmap project. *Am. J. Hum. Genet.* **84**(2), 251–258 (2009)
17. Google Genomics: (2015) <https://cloud.google.com/genomics/>
18. Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., Erlich, Y.: Identifying personal genomes by surname inference. *Science* **339**(6117), 321–324 (2013)
19. Hawkins, N.: The impact of human gene patents on genetic testing in the UK. *J. Gene Med.* **13**(4), pp. 320–324 (2011)
20. Hayden, E.C.: Privacy protections: the genome hacker. *Nature* **497**, 172–174 (2013)
21. Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W.: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4** (2008)
22. Huang, Z., Ayday, E., Hubaux, J.-P., Fellay, J., Juels, A.: Genoguard: protecting genomic data against brute-force attacks. In: Proceedings of IEEE Symposium on Security and Privacy (2015)
23. Humbert, M., Ayday, E., Hubaux, J.-P., Telenti, A.: Addressing the concerns of the Lacks family: quantification of kin genomic privacy. In: Proceeding of the 20th ACM Conference on Computer and Communications Security (CCS) (2013)
24. Humbert, M., Ayday, E., Hubaux, J.-P., Telenti, A.: Reconciling utility with privacy in genomics. In: Proceedings of ACM Workshop on Privacy in the Electronic Society (WPES) (2014)
25. Im, H.K., Gamazon, E.R., Nicolae, D.L., Cox, N.J.: On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. *Am. J. Hum. Genet.* **90**(4), 591–598 (2012)
26. Johnson, A., Shmatikov, V.: Privacy-preserving data exploration in genome-wide association studies. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1079–1087 (2013)
27. Juels, A., Ristenpart, T.: Honey encryption: security beyond the brute-force bound. In: Advances in Cryptology—EUROCRYPT, pp. 293–310 (2014)
28. Kamim, L., Bogdanov, D., Laur, S., Vilo, J.: A new way to protect privacy in large-scale genome-wide association studies. *Bioinformatics*. 2013 Apr 1;29(7):886–93
29. Kantarcioglu, M., Jiang, W., Liu, Y., Malin, B.: A cryptographic approach to securely share and query genomic sequences. *IEEE Trans. Inf. Technol. Biomed.* **12**(5), 606–617 (2008). doi: [10.1109/TITB.2007.908465](https://doi.org/10.1109/TITB.2007.908465)
30. Kschischang, F., Frey, B., Loeliger, H.A.: Factor graphs and the sum-product algorithm. *IEEE Trans. Inf. Theory* **47**, pp. 498–519 (2001)
31. Lin, Z., Owen, A.B., Altman, R.B.: Genomic research and human subject privacy. *Science* **305**(5681), 183 (2004)
32. Loukides, G., Gkoulalas-Divanis, A., Malin, B.: Anonymization of electronic medical records for validating genome-wide association studies. *PNAS* **107**(17), 7898–7903 (2010)
33. Malin, B.A., Sweeney, L.: How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *J. Biomed. Inform.* **37**(3), 179–192 (2004)
34. National Human Genome Research Institute: Intellectual Property and Genomics. (2015) <http://www.genome.gov/19016590>
35. Nyholt, D., Yu, C., Visscher, P.: On Jim Watson's APOE status: genetic information is hard to hide. *Eur. J. Hum. Genet.* **17**, 147–149 (2009)

36. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers, San Mateo (1988)
37. Popa, R.A., Li, F.H., Zeldovich, N.: An ideal-security protocol for order-preserving encoding. In: Proceedings of the 2013 IEEE Symposium on Security and Privacy (2013)
38. Raisaro, J.L., Ayday, E., McLaren, P., Telenti, A., Hubaux, J.P.: On a novel privacy-preserving framework for both personalized medicine and genetic association studies. In: Privacy-Aware Computational Genomics (PRIVAGEN) (2015)
39. Shamir, A.: How to share a secret. *Commun. ACM* **22**(11), 612–613 (1979)
40. Shih, W.: A branch and bound method for the multiconstraint zero-one knapsack problem. *J. Oper. Res. Soc.* **30**, 369–378 (1979)
41. Stajano, F., Bianchi, L., Liò, P., Korff, D.: Forensic genomics: kin privacy, driftnets and other open questions. In: Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society (2008)
42. Sweeney, L., Abu, A., Winn, J.: Identifying Participants in the Personal Genome Project by Name. Harvard University, Cambridge (2013)
43. Wang, R., Li, Y.F., Wang, X., Tang, H., Zhou, X.: Learning your identity and disease from research papers: information leaks in genome wide association study. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, pp. 534–544 (2009)
44. Yu, F., Fienberg, S.E., Slavkovic, A.B., Uhler, C.: Scalable privacy-preserving data sharing methodology for genome-wide association studies. *J. Biomed Inform.* 2014 Aug;50:133-41
45. Zhou, X., Peng, B., Li, Y.F., Chen, Y., Tang, H., Wang, X.: To release or not to release: evaluating information leaks in aggregate human-genome data. In: Proceedings of the 16th European Conference on Research in Computer Security (ESORICS'11), pp. 607–627 (2011)

Chapter 19

Encryption and Watermarking for medical Image Protection

Dalel Bouslimi and Gouenou Coatrieux

Abstract Among security mechanisms studied today, a growing interest is given to the protection of digital content especially with the help of encryption, watermarking and, more recently, of their combination. If encryption can be considered as an “*a priori*” protection mechanism since once data are decrypted they are no longer protected, it should be completed by an “*a posteriori*” protection, a protection watermarking can provide. Basically, when it is applied to images, watermarking modifies or modulates in an imperceptible way the image pixels’ gray level values in order to insert a message. This message can be used to assess the origins as well as the integrity of the image. As defined, watermarking provides an “*a posteriori*” protection as the image content is still available for interpretation while remaining protected. In order to benefit from the complementarity of these two mechanisms, in terms of *a priori/a posteriori* protection, different approaches have been proposed. In this chapter, focusing in the healthcare domain, we present the different purposes for which these approaches have been proposed for.

19.1 Introduction

The rapid evolution of multimedia and communication technologies has boosted the sharing and remote access to patient data. Care becomes ubiquitous and, at any time, at any place, it is possible for any citizen to access his or her medical data (see for example the French electronic patient record DMP¹ or the development of telemedicine applications [11]). For that purpose, the healthcare system infrastructure is adapting and quickly evolving (e.g., medical cloud computing [1], data warehouse [38]). Systems interconnect; data are exchanged and shared not

¹DMP: “Dossier Médical Personnel”, supported by The ASIP santé; dmp.gouv.fr.

D. Bouslimi (✉) • G. Coatrieux

Institut Mines-Telecom, Telecom Bretagne, Technopole Brest-Iroise, France

Laboratory of Medical Information Processing, LATIM INSERM UMR 1101, France

e-mail: Dalel.Bouslimi@telecom-bretagne.eu; Gouenou.Coatrieux@telecom-bretagne.eu

only for personalized patient care but also for epidemiological studies, economical evaluations and so on. If these technological advances offer significant benefits for medical providers and patients, it is obvious that they also increase security concerns. Taking as an example a telemedicine application where a practitioner outsources some Electronic Medical Records (EMR) of one patient, he or she loses the control on these data, data the practitioner is however responsible for and which have been given to him or her in complete trust by the patient. These concerns are more critical within public cloud computing frameworks, where a hospital can externalize some parts of its Information System (IS). It is important to recall that trust between a patient and health professionals is the key element of their relationship. Which confidence will you have in a practitioner or a healthcare establishment, the information system of which can be scanned from the Internet? (<http://www.cnil.fr/linstitution/actualite/article/article/cloture-de-la-mise-en-demeure-adoptee-a-lenccontre-du-centre-hospitalier-de-saint-malo/>, 2014). Today, ensuring security is a complex task that relies on the definition of a security policy. Such a policy expresses security objectives in terms of confidentiality, integrity, traceability and so on, and also establishes how it should be deployed based on different security mechanisms.

In this chapter, we focus on the protection of medical images by means of watermarking and cryptographic tools, the combination of which is very recent in healthcare. Medical imaging plays a central role today at almost all steps of healthcare, going from the diagnosis to the therapy follow-up or clinical research. As a consequence and as any other medical data, images are more and more distributed and shared supported for example by the constitution of regional or national Picture Archiving and Communications Systems (PACS), as in the US [34] and Canada (<http://www.infoway-inforoute.ca>).

Nowadays and in practice, protection of medical images is achieved by cryptographic means, especially encryption and digital signatures, as recommended by DICOM,² the medical imaging standard of reference. The kind of protection that encryption mechanisms offer is however mostly “*a priori*” in the sense that an image is protected until its decryption or, equivalently, before the access to its content is granted. Several questions arise then about the security of the image once decrypted. Here comes the interest for an “*a posteriori*” protection that allows the user accessing the image content while maintaining it protected—a kind of protection the watermarking technology is appropriated for. Basically, when it is applied to images, watermarking modifies or modulates the image pixels’ gray-level values in an imperceptible way in order to encode or insert a message (i.e., the watermark). This message can be used, for example, to verify the image integrity or for traceability purposes. The combination of cryptographic means and watermarking can therefore be used to achieve a continuous *a priori* and *a posteriori* protection and, as we will presented in the chapter, it provides solutions able to achieve different security objectives at the same time.

²DICOM: Digital Imaging and Communications in Medicine; <http://medical.nema.org>.

This chapter is divided into four sections. In Sect. 19.2, we summarize the security issues and needs one must consider when sharing medical data in general and in telemedicine applications. Then, in Sect. 19.3, we present encryption mechanisms and, in particular, the main ones the DICOM standard is interoperable with. Section 19.4 is devoted to watermarking in the context of medical imaging. In Sect. 19.5, we give an overview of different approaches that merge encryption and watermarking to enforce data protection. We also describe in detail one practical joint watermarking-encryption algorithm, which gives access to the outcomes of the image integrity and origin, even though the image is stored encrypted that is to say without giving access to the image content. As we will see, this method is transparent or compliant with the DICOM standard. More precisely, if a system is not watermarking interoperable, it can still decrypt the images and access its contents.

19.2 Security Needs for Medical Data

In this section, we first specify – in the broad sense – the security needs for medical information. These ones, imposed by the legislator through strict legislative and ethical rules, constitute a general framework that can be refined depending on the applicative context. This is the case for example of telemedicine applications we consider in the second part of this section and where traceability concerns are of importance, especially in case of litigation.

19.2.1 General Framework

In general, security is imposed by the legislator to protect citizens' rights. There are many sources of regulations, national and international, criminal and ethical, of general application (like those for software protection) or specifically devoted to medical data. Most of the time, the purpose of medical regulations is to create a relationship of trust between patients and health professionals. Indeed, a patient will be better taken in charge if he or she gives a view of his/her symptoms without any shadows. These regulations give rights to patients and duties to health professionals about the security of the data they possess. As example, in France, one has to respect the "Informatique et Liberté" Law of January 6th, 1978, completed by the Law of January 1st, 1994 (<http://www.cnil.fr/fileadmin/documents/approfondir/textes/CNIL-78-17definitive-annotee.pdf>), which establish security of digital data and, by extension, of medical data, as one right of citizens and consequently of patients. This Law requires that healthcare professionals take "all possible measures to safeguard information security and particularly to prevent its distortion, damage, or communication to any unauthorized third parties". Non-respecting this regulation is subject to criminal penalties.

From a practical and a technical point of view, committees for standardization can also give some practice recommendations on how to best protect medical information. Among them are Working groups, such as those put in place by the European Standards Committee TC/251 Medical Information (Working Group III [35]), IHE (Integrating the Healthcare Enterprise) (<http://www.ihe.net/>), HL7 (Health Level 7) (<http://www.hl7.org/>), and others. Based on all the current regulations and recommendations, the following should be ensured in order to meet the security requirements of medical digital content [4]:

Confidentiality: Confidentiality is the property which ensures that only authorized users in normal conditions have access to the information.

Integrity: A proof that a piece of information has not been modified in a non-authorized way.

Availability (of personal medical information): Ability of an information system to be used under the normal conditions of access. Its absence may be the consequence of a partial or total destruction of the data or a denial of access. The origins of the destruction may be different (physical or logical attack) and trigger immediate or delayed loss.

On the other hand, there is an intrinsic link between the quality of the information and the quality of care. L. Dussere et al. have defined in [41] the quality criteria or properties a piece of information should possess in order to be valid: Relevance (interest at one instant ...); Accuracy; Actuality (e.g., medical knowledge is in perpetual evolution, observations must be dated ...); Availability (dual to confidentiality); Exhaustiveness (any piece of information can be important ...); Fineness of the operator's judgment (all interpretations are operator-dependent); and, Reliability. As defined by L. Dussere et al., reliability corresponds to the confidence one can have on the information and, consequently, this property has a direct link with the concepts of integrity and authenticity. Indeed, any piece of medical information that is "sensitive" for patient care should not be modified nor altered (integrity). At the same time, because a medical record is most of the time nominative, its access is limited to authorized users (confidentiality), and a recipient as well as an issuer should be identified (authenticity).

Both considerations about security and quality of medical information can be merged, leading to the notion of data reliability [25]. Reliability relies on the outcomes of: (1) data integrity, and (2) data authenticity—a proof of the information origin, as well as of its attachment to one patient. If a piece of information is reliable, it can be used in complete trust by the physician.

These security objectives constitute a general framework for medical applications. However, they have to be refined depending on the application context. For instance, in the case of telemedicine applications, in addition to confidentiality, reliability and availability, there is a need for security functionalities able to support evidence, in case of litigation in order to establish responsibilities of each party involved in the communication.

19.2.2 Refining Security Needs in an Applicative Context: Telemedicine Applications as Illustrative Example

Supported by the development of multimedia and communication technologies, telemedicine applications enables distant patient care by offering to health professionals an easy access, management and processing of data. The development of telemedicine began more than 20 years ago and has since demonstrated its efficiency [78]. Even though technology strongly increased the quality and speed of image transmission, telemedicine only got a legal recognition in France in 2009 and its legal framework was defined in 2010 [5, 37]. In this context, when a patient suffers of a prejudice related to a diagnosis error, it is necessary to determine the respective liabilities of the practitioners involved in the diagnostic/therapeutic process. Professional negligence will be argued if one of the practitioners involved in the telemedicine has had a negligent attitude, i.e., falling short of what might have been expected from him/her regarding his/her field of competence. To assess this possible lack of professionalism, and, thus, possible legal liability of one or both involved practitioners (joint liability), several questions will arise to allow the judge to reach a conclusion: Who made the request? What was requested? When? For whom? What documents were provided/requested? Who answered? What? When? Regarding which documents?

To discover the process that originated the error, all the elements involved in the transaction must be carefully stored, with no means of modification (**Need 0 (N0)**). These elements are clearly defined in France by the Confidentiality decree of May 5th, 2010 (<http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000022932449&categorieLien=id>). Indeed, it is necessary that:

Need 1 (N1): All transmitted data have to be saved with the identity of all practitioners, the name of the patient, the date and the time of the transaction.

Need 2 (N2): The date, time and substance of the answer of the referent practitioner must be strongly linked to the documents that he/she received before sending it.

Need 3 (N3): The substance of the answer of the referent with the identifiers of the physician, the specialist, the date and the time of the transaction, must be saved.

Need 4 (N4): Both practitioners must be identified in such a way that they cannot repudiate their respective messages.

Need 5 (N5): All elements involved in the transaction must be stored, with no means of modification, and rendered unreadable from individuals with unauthorized access.

To sum up, in the case of telemedicine applications, in addition to the needs of confidentiality, reliability and availability required by the legislator in a general framework, it is also necessary to ensure data *traceability* and data *non-repudiation*. Data traceability aims at tracing data throughout its existence. Data non-repudiation guarantees that physicians cannot deny the emission or reception of data. So, in the deployment of a telemedicine application, all possible means must be implemented to ensure simultaneously these needs. To this end, an original solution can take

advantage of the complementarity of cryptographic tools and watermarking to offer continuous protection. When considered together, one offers an *a priori* protection (i.e., encryption) and the other an *a posteriori* protection, letting the user to access the information while maintaining it protected by watermarked security attributes for integrity, authenticity and traceability purposes. Before explaining how watermarking and encryption mechanisms have been combined in the research literature, we describe in what follows some encryption and watermarking primitives.

19.3 Encryption Mechanisms: An *A Priori* Protection

Encryption, or more generally, cryptography has received interest since antiquity. Many cryptosystems exist today, some of which are standardized and/or integrated within healthcare standards like DICOM. From a technical point of view, these cryptosystems can be differentiated according to various criteria, the classical ones being the symmetric/asymmetric character of a cipher or the way it works on a clear text: *block cipher* vs. *stream cipher*. In the following sections, we describe these categories and their differences, with a special emphasis on the symmetric block cipher and the stream cipher; cryptosystems that are better adapted to the protection of huge volumes of data.

19.3.1 Symmetric/Asymmetric Cryptosystems & DICOM

As illustrated in Fig. 19.1, the basic purpose of encryption is to transform a plaintext into a ciphertext in such a way that only the authorized parties can read it by means of a decryption process. The ciphertext is by definition unreadable or incomprehensible. According to the principles of modern cryptography, the security of a cryptosystem relies on the knowledge of its parameters, i.e., the encryption and decryption keys, and not on the knowledge of the applied encryption algorithm. Depending on the relationship between the encryption and decryption keys, one can distinguish two kinds of cryptosystems:

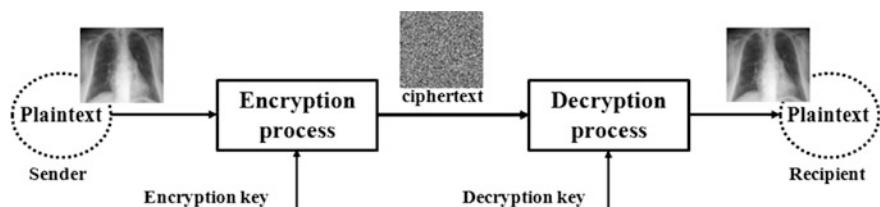


Fig. 19.1 A cryptosystem

- **Symmetric cryptosystems**, where encryption and decryption keys are identical. The DES (Data Encryption Standard) or the triple DES (DES applied three times consecutively), and more recently the AES (Advanced Encryption Standard) [33] all fall under this category [68].
- **Asymmetric cryptosystems**, like the RSA and Paillier [71], are controlled by a pair of keys that corresponds to one user. The asymmetry comes from the fact that keys are distinct but mathematically related and if a text is encrypted with one key, it can only be decrypted using the second key. In a pair, one key is said *private*, only known by the key pair owner and is kept secret, while the other key is said *public* and can be distributed to whomever needs it. Depending on the key that is used during the encryption process, different scenarios can be considered:
 - If the sender encrypts the message using the public key of the recipient, only the latter will be able to decrypt it with his/her private key. Data **confidentiality** is consequently ensured.
 - If the sender encrypts the message using his/her private key then any recipient will have to use the sender's public key to access the message. Message origin is assessed and as nobody else than the sender knows the private key, we have **non-repudiation**.
 - If one wants to ensure non-repudiation in addition to confidentiality, he or she has to encrypt two times the message: the first one using his or her private key and the second with the public-key of the recipient.

Symmetric cryptosystems are 100 to 1000 times faster than some asymmetric cryptosystems. They are thus more appropriate when huge volumes of data have to be protected. That is especially the case in medical imaging [91]. Notice that DICOM recommends the use of the DES, Triple DES, as well as of the AES algorithms to encrypt the image pixels and other pieces of information within the image file header.

Asymmetric cryptosystems are usually applied to initiate a secure communication between users or systems. As they offer non-repudiation in addition to confidentiality, users can authenticate each other and confidentially agree on the secret key of a symmetric cryptosystem, subsequently used to secure their communication. DICOM recommends the use of TLS (Transport Layer Security [39]) which is in part based on such an initialization protocol.

19.3.2 Block Cipher/Stream Cipher Algorithms

There exist two types of encryption algorithms: block cipher algorithms and stream cipher algorithms. Block cipher algorithms, like AES and DES, operate on large blocks of plaintext, whereas stream cipher algorithms such as RC4 or SEAL (Software-Optimized Encryption Algorithm) [86], manipulate stream of bits/bytes of plaintext. To highlight the differences of these two categories of algorithms, we

give in what follows an overview of AES, a block cipher algorithm recommended to be used by DICOM, and a well-known stream cipher algorithm, RC4 (Rivest Cipher 4), which is interesting because it can be easily combined with watermarking, as we will see in Sect. 19.5.

19.3.2.1 The AES Block Cipher Algorithm

The Advanced Encryption Standard (AES) manipulates blocks of 16, 24 or 32 bytes of plaintext. To encrypt a plaintext, it repeats a certain number of times a round; a round is a set of transformations. The round number depends on the encryption key size. Let us consider, for instance, a key size of 16 bytes which makes AES executing ten rounds. In this situation, a block B_i of 16 bytes too is re-organized in a 4×4 matrix before entering into the following process of AES (see Fig. 19.2):

1. “AddRoundkey”, where B_i is combined, through a *xor* operation, with a sub-key K_i derived from the encryption key using the Rijndael’s key schedule [33].
2. “N-1 Rounds”, each composed of the following consecutive steps:
 - a. “SubByte”, where each block of bytes is substituted with another block issued from a lookup table: the Rijndael S-box.
 - b. “SubByte”, where each row of the matrix B_i is right-shifted cyclically by different offsets.
 - c. “MixColumn”, where each column of the matrix B_i is treated as a polynomial multiplied in the Galois Field (GF) (28) with the following matrix (the reader must refer to [33] for more details):

$$\begin{pmatrix} 2 & 3 & 1 & 1 \\ 1 & 2 & 3 & 1 \\ 1 & 1 & 2 & 3 \\ 3 & 1 & 1 & 2 \end{pmatrix}$$

- d. “AddRoundkey”.
3. “Final round”, which is composed of the three following steps: “SubByte”, “ShiftRow” and “MixColumn”.

AES can be considered as secure since, up to date, it can only be broken by a brute force attack, which consists of testing all possible combinations of plaintext in order to find the encryption key. For instance, with this attack, 2^{128} operations are needed to recover an AES key of 16 bytes. Notice that the AIX (Adaptive Image Exchange), which is a HIPAA-compliant Medical Image Exchange that allows a practitioner or healthcare establishments to share medical images, uses the AES with 16-byte keys (<https://www.pacsclubevie.com/docs/aixhipaa.pdf>). This key size is also accepted by the DICOM standard, which however allows using longer keys. In [80], Snyder et al. recommend the use of AES with 32-byte keys in order to

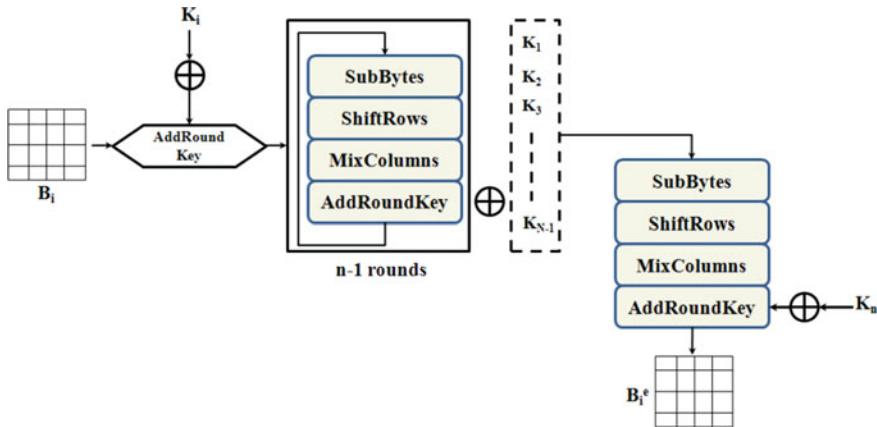


Fig. 19.2 General scheme of the AES algorithm

ensure a good compromise in-between the level of confidentiality and computation time. Indeed, data encryption induces delays in data access and consequently on the hospitals' workflow.

Block Cipher Modes of Operation The concept of mode of operation refers to the manner in which blocks of plaintext (a sequence of bytes) are treated at the encryption stage (*resp.* decryption stage). There exist many modes of operation: Electronic Codebook (ECB), Cipher Block Chaining (CBC), Output Feedback (OFB), Cipher Feedback (CFB), Counter (CTR), etc. (see [86]). The CBC mode of operation is recommended by DICOM when using AES. As depicted in Fig. 19.3, when the CBC mode is activated the plaintext block is combined with the previous ciphertext block through a *xor* operation, before being encrypted with the above AES process. If we denote B_i^e the encrypted version of a block B_i and B_{i-1}^e is the previously encrypted block, B_i^e is given by: $B_i^e = AES(B_i \oplus B_{i-1}^e)$.

19.3.2.2 The RC4 Stream Cipher Algorithm

General Principles of Stream Ciphers The basic way a stream cipher algorithm works is described in Fig. 19.4. It combines the bits/bytes of plaintext $T = [t_1, \dots, t_i, \dots, t_n]$ with a secret keystream of bits/bytes $K = [k_1, \dots, k_i, \dots, k_n]$ issued from a pseudo random number generator (PRNG), through a *xor* operation, typically. The keystream generation depends on a secret key K_e . Thus, bits/bytes of cipher text $C = [c_1, \dots, c_i, \dots, c_n]$ are usually defined as: $c_i = t_i \oplus k_i$

The specificity of such stream cipher algorithms resides in how the bit/byte keystream is generated by the PRNG (the reader should refer to [86] for the description of well-known stream ciphers' PRNGs). Notice also that although the *xor*-operator is most frequently used, some have instead proposed to use the addition

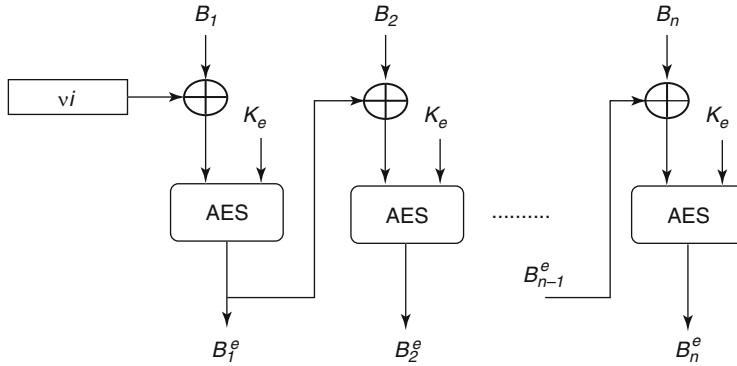


Fig. 19.3 AES encryption in CBC mode. B_i , B_i^e and K_e denote the plaintext block, the encrypted block and the encryption key, respectively. iv is a random initialization vector

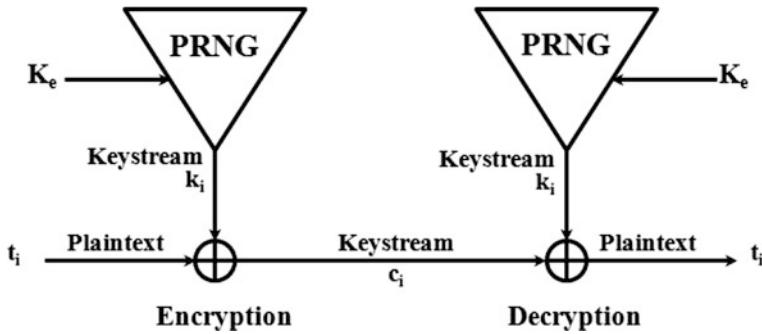


Fig. 19.4 Encryption/decryption processes of a stream cipher algorithm

modulo- d operation (i.e., $c_i = (t_i + k_i) \text{mod}(2^d - 1)$), where d is the image depth) [92] in order to be able to use the arithmetic operations.

Some of the main advantages of this type of algorithms are that they are simple and operate at a higher speed than block cipher algorithms. However, we should note that the latter algorithms are more secure due to the complexity of the transformations they are based on.

Presentation of the RC4 Algorithm RC4 is a very popular stream cipher algorithm. It is used in standard security protocols like TLS [37], WEP (Wired Equivalence Privacy [57]) and WPA (Wi-Fi Protocol Access [6]). The algorithm relies on the *xor*-operation between the plaintext and the keystream issued by the RC4 PRNG. The latter is based on two steps:

- “Initialization”, where a table of 256 bytes is filled by repeating the encryption key as often as necessary.
- “Byte keystream generation”, where the elements of the table are combined by applying permutations and additions to generate the keystream.

19.4 Watermarking: An *A Posteriori* Protection Mechanism

19.4.1 Principles, Properties and Applications

Watermarking [31, 32, 55, 87] embraces the principles of information hiding, which also includes steganography [32, 46, 55, 87], a discipline as old as cryptography, and uses them to establish secret communication. By definition, information hiding allows the insertion of a message within a host document by modifying it in an imperceptible way. When it is applied to an image (host signal), the pixels' gray-level values of the image are modified or modulated so as to encode a message. If for steganography there is no a priori link between the message and the host content, its objective being at first to secretly communicate the message, this is not the case for watermarking. Indeed, watermarking purposes find their origin in the 90s when, with the Internet advances, where questions arose about intellectual property protection of digital data: how to protect digital content that can be recopied without loss of quality? How to be certain that a buyer will not illegally redistribute his or her copy? . . . Questions that can be answered by embedding or watermarking the buyer or owner identifier. Nevertheless, the main interest of watermarking is that it leaves access to the data, while maintaining them protected by the watermark: watermarking is an *a posteriori* protection mechanism. Since then, watermarking has been proposed for other applications as well as in other domains, like in healthcare, where the protection of intellectual property is of less concern. Different watermarking applications are foreseen in healthcare [26], which depend on the relation in between the message and its host, as discussed below.

Image Authenticity Control Image authenticity control can be ensured by inserting a proof of the image origin and of its attachment to one patient. The underlying question is how to build a relevant “authenticity code”. Herein, DICOM can conveniently be exploited as it imposes the creation of a Unique Identifier (UID) for any acquired images. The UID refers to the machine, health institution, patient, time and date of the examination, etc. Consequently, one can get a simple authenticity code by combining such a UID with a patient identifier. It becomes therefore possible to verify that the UID and the patient identifier present in the header DICOM file are identical to those embedded into the image. Notice that patient identifiers are sometimes defined in a legislative way [78]. In France, for example, a National health identifier exists since 2007 (Article L1111-8-1 of the public health code [57]). Such identifiers must be used with care as they may induce privacy issues. Indeed, they reveal the exact patient identity. Beyond that, an authenticity code can be used to establish secure links between Electronic Medical Records of one patient [27]. For instance, images can be watermarked with the unique identifier of the medical report they are related with. From the image pixels, it will be possible to know that the images have been already analyzed, while making possible to retrieve the report.

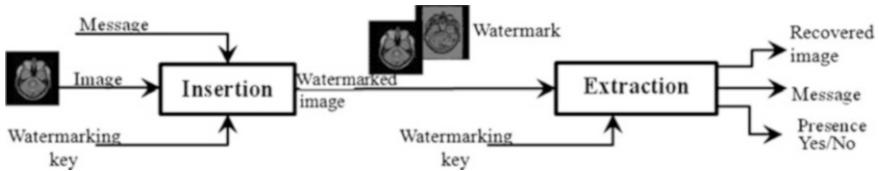


Fig. 19.5 The principle of watermarking chain

Image Integrity Control Image integrity control based on watermarking of a proof, allows verifying that the host data have not been altered or modified by non-authorized persons. In the case of medical images, one common solution consists in embedding a digital signature or a cryptographic hash [86] computed over the image or some parts of it [61, 72]. We recall that a cryptographic hash function is an algorithm which transforms a message of arbitrary size into a hash of fixed and small size (e.g., 160 bits in the case of the SHA-1—Secure Hash Algorithm [42, 86]), while inter alia having good dispersion properties (two similar documents have very distinct hashes). Instead of using a cryptographic hash which is very sensitive to any host modifications, it is possible to embed a perceptual hash [86] that will raise an alarm only for some unauthorized modifications. In [29], the authors embed a set of image signatures (cryptographic hashes and an image moment signature) that allow them to: detect an image modification occurred; localize the image tamper and by complementarity indicate which parts of the image can trustfully be used; and identify the nature of the modification. Some others [22, 64] have proposed to extend in healthcare the concept of self-correcting images introduced by Fridrich and Goljan [47]. The basic idea is to embed into the image redundant data that allow reconstructing the image content after it has been modified.

Metadata Insertion The purpose of this application is to insert watermarked metadata that can facilitate the use of the host content or its management, or to make it more informative. Very few studies have been conducted on this theme in healthcare [28, 65]. In [75], the record indexing information is embedded into the media content to identify the media. This makes possible to repair damaged index tables of a relational database by completing missing pieces of information from the watermarked images or data. Beyond these examples where database coherence and safety can be improved, one can also offer new services. In [28], within the framework of an e-learning application, images are watermarked with an “image knowledge digest” which semantically describes the image content while providing rules for comparing images in terms of diagnosis or lesion.

19.4.1.1 A General Chain of Watermarking

A classical watermarking scheme is composed of two steps, as shown in Fig. 19.5:

- *Watermark insertion process*, which allows inserting a message within an image by modulating its pixels or some transformed coefficients (e.g., DCT—Discrete

Cosine Transform [81], TFD—Discrete Fourier Transform [90]). Note that the “distortion” introduced or the signal of difference between the original image and its watermarked version constitutes the watermark. This watermarking process usually depends on a secret watermarking key which can be used, for instance, to secretly select pixels or coefficients to watermark. There exist several types of watermarking modulations, some of which are described in Sect. 19.4.2.

- **Watermark detection/extraction process.** Detection is a prerequisite to message extraction, but both processes can be blind or non-blind [49, 51] depending on the need to have or not access to the original image. It is obvious that methods requiring accessing the original image at the detection process are limited in terms of applications. In between blind and non-blind schemes we can find semi-public schemes [31], where the watermarking key contains some pieces of information about the original image (such as a summary of it); pieces that will help the decoder to extract the message.

19.4.1.2 Basic Properties of a Watermarking Algorithm

Watermarking algorithms are characterized by a set of different properties in-between which a tradeoff has to be established:

Robustness: A watermark is considered as robust if after a transformation or an “attack” of the watermarked image, the watermark is still available. Attacks are categorized as either malevolent or innocent. By definition, an *innocent attack* is a process foreseen in the exploitation cycle of images. It corresponds generally to processing operations, like filtering and lossy compression (e.g., JPEG—Joint Photographic Experts Group). On the contrary, a *malevolent attack* aims at removing or modifying the watermark for an illegal use of the image.

Capacity of insertion: It corresponds to the amount of information that one can embed into an image and is usually expressed in bpp (bits of message per pixel of image), i.e., the number of bits of message that is embedded per image pixel.

Reversibility: This property ensures that once the watermark is detected and extracted, it can be removed from the image, i.e., inverting the distortions introduced during the insertion process, allowing the recovery of the original image. Such a property is often desired in applications where the quality of documents is a strong constraint, as in healthcare.

Security: The basic idea behind this property is that the watermark should be very difficult to counterfeit without the knowledge of the watermarking key as for cryptography. Notice that a robust watermarking scheme is not necessarily secure [50]. To enhance message security, some propose to asymmetrically encrypt it before its embedding [80]. So, even though a “pirate” may extract the information, he or she cannot access the clear message or generate a valid message without knowledge of the encryption keys.

Imperceptibility: By definition a watermark should be invisible to the user. For medical images, this notion of “imperceptibility” is important. Indeed, modifying

the gray-level values of a medical image may induce risks of misinterpretation. The measure of image degradation today revolves around objective measures like the Peak Signal to Noise Ratio (PSNR), the Signal to Noise Ratio (SNR) [56], and the Normalised Root-Mean-Square Error (NRMSE) [43]. Although these solutions are not always representative, their implementation is of less complexity and of less cost compared to a subjective evaluation. The latter requires asking a group of users whether they can distinguish watermarked images from the original ones.

The choice of a watermarking method depends on the application framework, striking a compromise between different requirements. Needs of a telemedicine application may differ from those inside the hospital. In the former case the information is outsourced, with strong traceability issues, while in the latter case, we have a better idea of who may access the data (see [25] for more details).

19.4.2 Watermarking Medical Images

In many aspects, imperceptibility is the main constraint in healthcare. Medical tradition is very strict with the quality of biomedical images, in that it is often not allowed to alter in any way the bit field representing the image (non-destructive). Lesions or findings can be very subtle and may be masked by distortions induced by the watermark. At the same time, it should be considered that the original captured image often undergoes certain processing, like enhancement and contrast stretching, windowing with parameters' values varying from one user to another [26]. Consequently, the watermark may become more or less “visible”. In [25], Coatrieux et al. propose to differentiate three kinds of watermarking schemes, each introducing its own constraints and limitations:

1. **Lossless or reversible watermarking:** Schemes in which the watermark can be removed with the exact recovery of the original image.
2. **Region of Non Interest (RONI) watermarking:** This scheme is based on the definition of some regions of interest (ROIs), to be left intact, and regions of insertion. Because it is assumed that ROIs contain all relevant information for the diagnosis, one can consequently revert to methods with higher capacity and robustness, and has to ensure that the watermark in RONIs is related to ROIs.
3. **All other methods:** We herein refer to insertion methods that use the whole image, and bring about imperceptible alterations in the pixels, as commonly used in all multimedia signals. Nevertheless, a psychovisual model [59, 94, 98], like those used for natural images so as to adapt the watermark amplitude locally into the image, does not yet exist for medical imaging. However, if we consider that in some applications like telemedicine the use of lossy image compression is accepted, we can assume without too much risk that classical watermarking schemes can be applied even though they bring different distortions. Notice that different studies have shown that under some constraints lossy compression can

be exploited [89]. Similar studies have however to be conducted for medical images. In [75], robust watermarking has been considered after a physician has selected the maximum power of the watermark just under the level of interference with the patient diagnosis.

19.4.2.1 Basic Lossy Watermarking Modulations

Watermarking techniques can be classified in many ways. It is however habitual to differentiate additive from substitutive modulations.

Additive Methods Starting from a message (e.g., a sequence of bits), an additive method generates a signal, a *watermark* which is added to the image or a transformation of it (DCT [81], WT [8, 19, 66], etc.) Spread-spectrum based watermarking [31, 32] is a classical example of such a kind of technique. A simplified view of it is the following one (see Fig. 6). It embeds one bit b_i of a message by adding to the image I a pseudo-random sequence w_i multiplied by $d_i = 1 - 2b_i$. As defined, the watermarked image I_w is given as: $I_w = I + \alpha d_i w_i$, where α controls the watermark strength or *amplitude* and thus allows controlling the watermark robustness-imperceptibility compromise. α can be locally computed based on psychovisual masking [32]. The embedding of a message of N bits adds to the image the watermark: $W = \sum_{i=1}^N \alpha d_i w_i$. The presence of this watermark, as well as the message extraction, is usually based on correlation techniques, which implies the orthogonal nature of the pseudo-random sequences w_i . The sign of each correlation product $\langle W, w_i \rangle$ gives the value of the embedded bit b_i .

Substitutive Methods In order to embed one symbol of a message, such a modulation replaces one piece of information of an image, a feature or a characteristic of it, with another that is issued from a dictionary that encodes the desired symbol value. As a consequence, message detection and extraction simply stands in an image feature re-reading or interpretation. Some well known substitutive watermarking modulations are the Least Significant Bit (LSB) substitution and the Quantization Index Modulation (QIM).

A. Least Significant Bit (LSB) Substitution Modulation

This algorithm is the simplest substitutive modulation [9, 24]. It replaces or substitutes the least significant bits (LSBs) of the image pixels by those of the message (see Fig. 7). At the extraction stage, one just has to interpret the parity of watermarked pixels' values so as to read/extract the message. If we denote by P_i^w the watermarked pixel and b_i is one bit of the message, then $b_i = LSB(P_i^w)$, where $LSB(\cdot)$ is a function that provides the LSB of the corresponding value. This method provides high watermark capacity with a low image distortion: not more than one gray-value per pixel. At the same time, it is very fragile as any modification of the image will destroy the message.

B. Quantization Index Modulation (QIM)

Due to its robustness and simplicity, QIM is today one of the most popular substitutive watermarking techniques for images [92] as well as for videos [88]. It has been also proposed for relational databases [45]. QIM relies on quantifying an image component (e.g., group of pixels or of transformed coefficients) according to a set of quantizers based on codebooks so as to insert a message. Basically, to each message m_{si} issued from a finite set of q_s possible messages $M_s = \{m_{si}\}_{i=0,\dots,q_s}$, the QIM associates the elements of a codebook $C_{m_{si}}$ such as:

$$C_{m_{si}} \cap C_{m_{sj}} = \emptyset, i \neq j \quad (19.1)$$

Substituting one component X of the image by its nearest element X_w in the codebook $C_{m_{si}}$ thus allows the insertion of m_{si} . This process can be noted as:

$$X_w = Q_{m_{si}}(X) \quad (19.2)$$

where $Q_{m_{si}}(X)$ is a function that determines the nearest element X_w of X within $C_{m_{si}}$. The distance between X_w and X corresponds to the watermarking distortion.

During the extraction step, the knowledge of the cell to which the received version of X_w , noted as Z , belongs is enough to identify the embedded message. The message \widehat{m}_{si} detected at the reception is then given by:

$$\widehat{m}_{si} = \arg \min_{i=1,\dots,p} |Z - Q_{m_{si}}(Z)| \quad (19.3)$$

Example—Let us consider one image component, such as a vector of pixels $X \in \mathbb{N}^N$, while dividing the \mathbb{N}^N dimensional space into non-overlapping cells of equal size. Let us also assume the insertion of a binary message, i.e., $m_{si} \in \{0, 1\}$; two codebooks C_0 and C_1 will thus be defined. To satisfy [Eq. (19.1)], each cell is associated to a codebook $C_{m_{si}}$, $i = 0, \dots, q_s$. As a consequence, one message m_{si} has several representations in \mathbb{N}^N . The insertion process is conducted as follows. If X belongs to the cell which encodes the message to be inserted, X_w (the watermarked version of X) corresponds then to the center of this cell; otherwise, X is moved to the center of the nearest cell that encodes the desired message. In the case that X is a pixel value, i.e., $X = x$ and $X \in \mathbb{N}$, C_0 and C_1 can be built up by uniformly quantizing x with a quantization step Δ , as illustrated in Fig. 19.6. In this example, cells centered on crosses represent $C_1(m_{si} = 1)$, whereas cells centered on circles represent $C_0(m_{si} = 0)$. Thus, x will quantize to the nearest cross or circle in order to encode m_{si} .

Nowadays, Dither Modulation (DM) [20] is probably the most common implementation of the QIM, due to its simplicity based on uniform quantizers. This scheme consists in shifting the host signal by a value (a dither) pseudo randomly

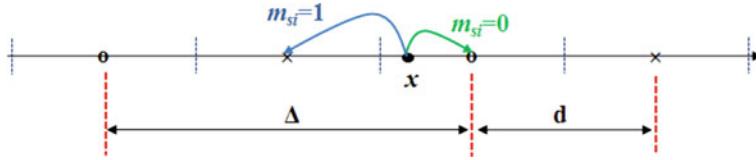


Fig. 19.6 Example of two codebooks' cells in the mono-dimensional space (i.e. x is a scalar value) considering an uniform quantization of quantization step Δ . Symbols o and x denote cells' centers that encode 0 and 1, respectively. $d = \Delta/2$ establishes the measure of robustness to signal perturbations

generated, known at the insertion and the extraction and belonging to the interval $[-\Delta/2, +\Delta/2]$, where Δ is the quantization step. With this modulation, the insertion process can be expressed as follows:

$$X_w = Q(m_{si})(X + d(m_{si})) - d(m_{si}) \quad (19.4)$$

where $d(m_{si})$ is the dither associated to the message m_{si} .

In order to enhance the robustness/distortion tradeoff, an improved version of QIM has been proposed by Chen and Wornell [20]. This version, known as Distortion Compensated QIM (DC-QIM), consists in re-injecting a fraction of the error quantification to the quantified signal. However, both QIM and DC-QIM are not robust against valuemetric scaling changes, which occur when modifying the image brightness, for example. To reduce this weakness, several extensions of the QIM were proposed. Among them are Angle QIM (AQIM) [21] and Rational Dither Modulation (RDM) [74]. As illustrated in Fig. 19.7, for message embedding, AQIM modulates or quantifies the polar representation angle of the vector representing the host signal rather than its modulus. It is therefore more robust than QIM and DC-QIM to the valuemetric scaling attack (VSA), which modifies first the signal modulus before its phase. With RDM, the quantization step is variable and is computed based on the previous watermarked samples.

RDM has been extended into the Perceptual-QIM (P-QIM) algorithm by Li et al. [63]. The originality of this algorithm relies on the use of a modified version of the Watson's perceptually model [94] in order to ensure the invariance to VSA, while taking into account the watermark psychovisual impact.

19.4.2.2 Lossless Watermarking

Since the introduction of the concept by Mintzer et al. in 1997 [69], several methods have been proposed for lossless watermarking. As above, we propose to distinguish them into two main classes: additive methods [53, 70, 84] and substitutive methods [16, 48]. In the case of *additive insertion*, the message m is first transformed into a watermark signal W next added to the host image I , leading to the watermarked image I_w . By doing so, and without paying attention, the watermarked image may

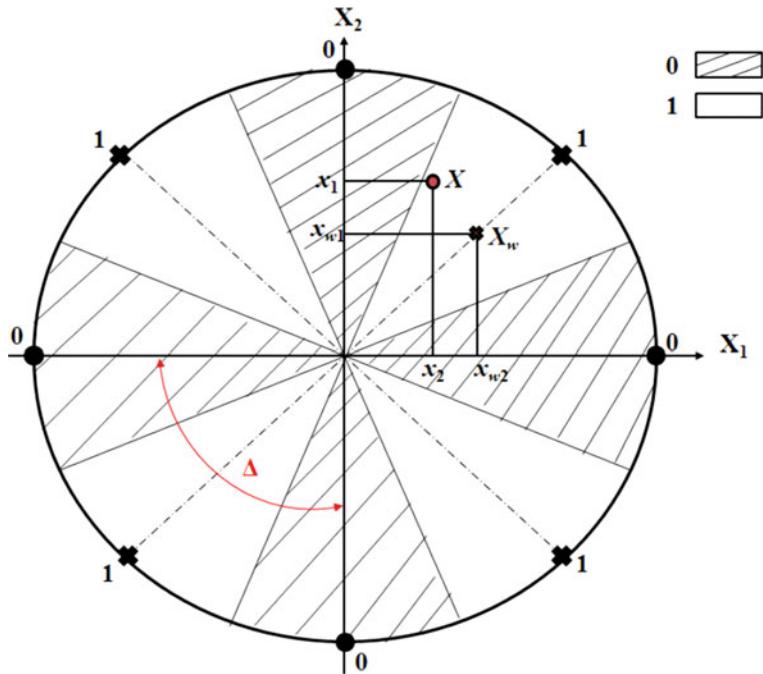


Fig. 19.7 Insertion of a binary message using AQIM. X_w represents the vector after the insertion of a bit equals to “1” within a host signal X associated to a vector in the N -dimensional space if N pixels constitute X . Δ is the quantization step, and circles and crosses represent the centers of the cells that encode the bits “0” and “1”, respectively

include pixels with gray-values outside the allowable image dynamic range (e.g., $[0 \dots 2^d - 1]$ for a d -bit depth image with positive integer values), introducing underflows (negative values) or overflows (values greater than $2^d - 1$). This is the main issue to be solved in order to achieve the reversibility property.

One first solution, suggested by Honsinger et al. [53], consists in using arithmetic modulo, that is to say: $I_w = (I + W)mod(2^d - 1)$. Reversibility is guaranteed by: $I = (I_w - W)mod(2^d - 1)$. The main drawback of this solution is a salt and pepper noise due to jumps between congruent values of the image histogram. An improvement to this method, proposed by Fridrich et al. [48] and De Vleeschouwer et al. [36], attenuates visual distortion by using arithmetic modulo with shorter histogram cycles or, more specifically, by splitting the signal dynamic in ranges of short size. However, the interest of these methods remains limited due to their low performance in terms of capacity.

An alternative to these modulations refers to *histogram shifting* (HS), introduced by Ni et al. [70]. As illustrated in Fig. 19.8, HS shifts a range of the signal histogram with a fixed magnitude Δ to create a “gap” near the histogram maxima. Samples of the signal with values associated to the class of the histogram maxima, the “carriers”, are then shifted to the gap or kept unchanged to encode “0” or “1”.

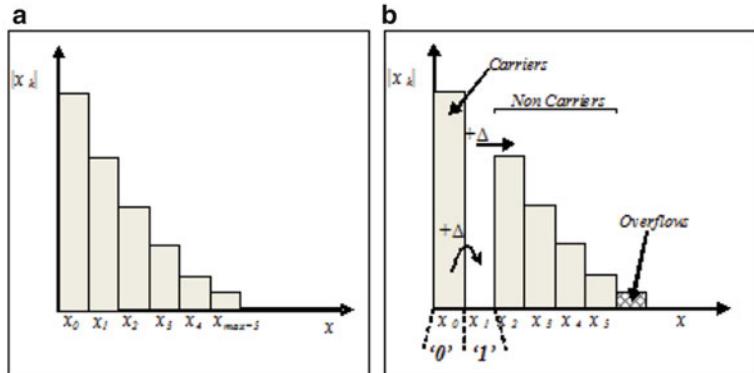


Fig. 19.8 Basic principle of the histogram shifting modulation: (a) original histogram, and (b) histogram of the watermarked data

Obviously, in order to restore exactly the original signal, the watermark reader needs to be informed of the positions of samples of values that would have been shifted out of the dynamic range (overflows in the example of Fig. 19.8). This requires the embedding of an overhead along with the message, which reduces the watermark capacity. It is clear that if the histogram is uniform, it is not possible to watermark the signal.

Since the work of Ni et al. [70], different refinements have been proposed. Instead of working in the spatial domain, several schemes apply HS to some transformed coefficients [96] or pixel prediction-errors [84]. Their histograms are concentrated around one single class maxima located at zero. This maximizes HS capacity and also simplifies the re-identification of the histogram classes of maximum cardinality at the reading stage. It also appears, from the most performant schemes at the moment [54, 84], that working with prediction-errors is the most appropriate way, histograms are of smaller variance.

Substitutive methods can be classified into two categories: Lossless Compression Embedding (LCE) techniques and Expansion Embedding (EE) techniques. Fridrich et al. developed one of the first LCE techniques in [48]. They losslessly compressed a bit-plane of the original image for data embedding. Since then, Xuan et al. have extended this approach to the wavelet domain [95], while Celik et al. [16] developed a generalized LSB substitutive technique.

Unlike LCE techniques, EE techniques expand the dynamic of the signal by shifting to the left the binary representation of a signal sample h , creating thus a new virtual least significant bit (LSB) that can be used for data insertion: $h_w = 2h + b$, where h_w is the watermarked version of h and b is one bit of the message. EE is combined with LSB Substitution applied to non-expandable signal samples, i.e., samples that cannot be expanded because of the limited dynamic of the signal or because of distortion constraints. Thus, one just has to read samples' LSBs to read the message. Again some reconstruction overhead needs to be embedded to

store substituted original LSBs, as well as the position of expanded samples. EE was experimented on: the difference between two adjacent pixels [93], vectors of adjacent pixels [3], wavelets coefficients of pixel blocks [62], etc.

Nowadays, most methods take advantage of HS and EE modulations at the same time [54, 84]. To do so, EE modulation is applied instead of simply shifting by a fixed amplitude samples of the carrier-classes. More clearly, HS is first used to create a gap in the pixel predicted error histogram, while EE is used to encode data in created gap. The capacity is the same as with HS but the distortion is minimized. The scheme presented in [29] is better than any of these solutions even though it only uses HS and can thus be improved if combined with EE. The originality of this scheme is twofold, it is based on: a “Dynamic Histogram Shifting Modulation”, which adaptively takes care of the local specificities of the image content and allows message embedding in textured areas where other methods fail to do; and on a signal sample classification, which purpose is to identify risks of underflows and overflows [28] and the most suited lossless modulation locally into the image [73].

19.5 Combining Encryption with Watermarking

Different approaches have been proposed to take advantage of the complementarity of encryption and watermarking mechanisms. Their basic objective is to ensure data confidentiality through encryption, while adding new security functionalities by watermarking. These new functionalities can be made available either in the spatial domain, i.e., after data have been decrypted, and/or in the encrypted domain. Actually, if most of these solutions focus on copyright protection, they can find an important place in the healthcare domain. In this section, we provide a brief overview of these solutions and further introduce a joint watermarking-encryption strategy, we originally proposed in the healthcare domain [12, 13] but not limited to.

19.5.1 *Continuous Protection with Various Security Objectives: A State of the Art*

Technically, depending on the way that watermarking and encryption are merged, we suggest to discriminate four main categories of methods:

1. **Watermarking Followed by Encryption (WFE):** In this case the host signal is first watermarked and then encrypted.
2. **Encryption Followed by Watermarking (EFW):** In this case a watermark is embedded into the host signal after its encryption.
3. **Commutative Encryption and Watermarking (CEW):** This case regroups techniques that simultaneously belong to the first two categories. More clearly, watermarking and encryption operations are independent and commutative, in

the sense that the host signal can be watermarked then encrypted or encrypted then watermarked. The result, i.e., the watermarked and encrypted signal, is the same in both cases.

4. **Joint Watermarking-Decryption (JWD):** In this case a watermark is embedded during the decryption process.

Beyond the aforementioned categories, there is a fifth one that we have recently introduced in [12, 13] and we call **Joint Watermarking-Encryption (JWE)**; it corresponds to the case where a watermark is embedded during the encryption process. We depict one implementation of such a scheme in Sect. 19.5.2 and we illustrate how it can be used to allow a user to verify the reliability of a document being encrypted or not.

19.5.1.1 Watermarking Followed by Encryption

This category regroups the first methods proposed in the literature. As depicted in Fig. 11, they simply consist of first embedding a watermark within the host data and then encrypting the data [60]. The embedded message can only be extracted after decryption. The main application for which these methods have been proposed is copyright protection. That is why in general, the watermark is embedded in a robust manner [60].

Nevertheless, a few applications have been devoted to medical data. In [83], the image is enriched with the insertion of patient data before being ciphered. In [79, 80], U. R. Acharya et al. propose to improve the security of patient records, some text or graphical signals, that are encrypted and then inserted into one image of the patient. Embedding is conducted by using LSB substitution in the spatial domain [79] or in the discrete cosine transform domain in the case that JPEG compressed images are considered [80].

19.5.1.2 Encryption Followed by Watermarking

Herein, the pursued objective is to watermark host signals in an encrypted form (see Fig. 12). These methods can be distinguished depending on if the watermarking modulation used is reversible or not. As for the previous class of methods, copyright protection constitutes the main application of these methods.

Non-reversible Watermarking of Encrypted Data The majority of these methods are based on *homomorphic encryption*, where the host signal is entirely encrypted using a homomorphic cryptosystem [18, 67, 82]. Let us consider $M = M_i$, where $i = 1, \dots, n$, to be the set of plaintexts and $C = C_i$, $i = 1, \dots, m$, to be the set of ciphertexts. An encryption algorithm E is said *homomorphic* if it satisfies the following relationship [44]:

$$\forall M_i, M_j \in M, E(M_i \perp_s M_j) = E(M_i) \perp_e E(M_j) \quad (19.5)$$

where \perp_s and \perp_e represent operations in M and C , respectively. \perp_e corresponds in general to the multiplication operation, while \perp_s can be an operation of addition, multiplication, *xor*, etc. Specifically, based on the homomorphic property, an operation in the encrypted domain or, more clearly, on the ciphertext (e.g., ‘ \times ’, ‘ $+$ ’, ...) is translated into another operation on the plaintext (e.g., ‘ $+$ ’, ‘ \times ’, ...). It becomes then possible to modify an encrypted image and to insert a watermark.

However, with this kind of approach, the embedded watermark is only available in the spatial domain. Indeed, the watermark is inserted in an encrypted form into the encrypted file and implies the detection of a “decrypted” watermark into the clear image. The approach described in [76] is another strategy which focuses on the authentication of the image’s origin. It consists in embedding in the encrypted gray-levels of the image the secret key used for its encryption. This key is encrypted twice with RSA: the first time with the private key of the sender (for authenticity), and the second time with the public key of the recipient (confidentiality). Here, a stream cipher algorithm is used to encrypt the image. At the reception, the encrypted key is extracted from the encrypted image which is next decrypted.

Reversible Watermarking of Encrypted Data These methods allows watermarking an encrypted signal in a reversible manner. The embedded watermark is removed from the encrypted data before the decryption process. Methods have been proposed to exploit the statistical properties of the encrypted and of the original signal in order to read and remove the watermark.

For example, in [77], Puech et al. insert a message into the encrypted image by substituting one bit of a pixel block with one of the message. At the recipient side, each block of the image is decrypted twice, testing the two possible values of the modified bit (0/1). The position of the watermarked bit must be known. The decrypted block with the minimum standard deviation is the original block. Indeed, a false bit generally prevents the correct document decryption, which appears as a noise. Another example of such an approach is given in [97], where a lossless compression is applied onto the encrypted image LSBs. The space gain is then used for message insertion. Again, inserted data are only available in the encrypted domain and need to be removed before or during the image decryption process.

19.5.1.3 Commutative Encryption and Watermarking

These techniques are often based on partial encryption [52] or invariant encryption [85]. In the latter case, the host signal is encrypted so that some of its features remain invariant to the encryption and decryption processes. These features are used for watermark insertion. For example, in [85], the image is encrypted by means of a random permutation of pixel positions without changing their gray values. The latter are watermarked using a histogram-based watermarking scheme before or after the encryption process [23]. Nevertheless, these “invariant” or “non encrypted” features can be exploited for an attack.

With such an approach, the embedded information is thus available in both encrypted and spatial domains. With partial encryption, only some parts of the host signal are encrypted, e.g., some Most Significant Bit (MSB) planes, while the rest is watermarked [10, 14, 15]. For instance, in [14, 15], the image is firstly decomposed into levels using the Haar wavelet transform. All coefficients are then transformed into bits and arranged in the form of bit planes. The Most Significant Bit planes of the image are encrypted using the AES, and the Least Significant Bit planes are watermarked using the QIM. In practice, these techniques are interesting in applications where large volumes of data (e.g., image or videos databases) are treated.

19.5.1.4 Joint Watermarking-Decryption

Joint watermarking-decryption systems have been proposed in the framework of video-on-demand for identifying a posteriori or tracing the buyer. They consist in embedding the fingerprint (the buyer's identifier) at the client side or more precisely during the decryption process. The objective is to reduce time computation and complexity on the server side (see Fig. 13). Basically, the same encrypted version of an original document is sent to all clients. At the client side, the decryption and watermarking operations are conducted simultaneously using the client decryption key; a key that is different from one client to another and which will induce differences in between clients' documents. These differences correspond to the watermark, that is to say to the client fingerprint, which uniquely identifies him or her.

The first scheme of this category of methods was proposed by Anderson et al. [7] in 1997 and is known as Chameleon, the principle of which is as follows. Let us consider the plaintext or clear text is constituted of bytes (e.g., grayscale pixels). During the encryption operation, four entries of a random look up table (LUT) are selected and then *xor*-ed with a plaintext component to obtain the ciphertext. The decryption process is similar to that of encryption, except that the decryption LUT is different. This one is derived from the encryption LUT by injecting errors in some entries. As a consequence, the deciphered text will be slightly different from the original text. To enhance the robustness against the collusion attack, i.e., when several clients combine their copies (e.g., by making their average), different extensions of Chameleon have been proposed. Among them are the fingercasting, proposed by Adelsbach et al. [2], and the solution suggested by Celik et al. [17] which operates with arithmetic operations and LUTs composed of real numbers.

The majority of methods that combine watermarking and encryption are intended to protect copyright. Because of their need for robustness, these methods introduce a strong distortion in the host signal. As discussed in Sect. 19.4.2, these methods may not be appropriate for medical images. Even though some are reversible, their poor scalability is also an inconvenience. Practitioners want quick access to their data. Beyond that, most of these methods give access to the embedded watermark only in one domain: encrypted or spatial domain. In some scenarios, it is nevertheless

interesting to access the watermark in both domains. This, for example, allows a user to verify the reliability of an image and to determine to which patient the image is associated to, without decrypting it. Nowadays, few solutions give access to the embedded watermark in both domains. They are essentially based on partial and invariant encryption techniques. Because these techniques do not encrypt all the data, they are faster. But at the same time, one may question their security. A better response is achieved if data are fully encrypted while giving access to watermarking functionalities, images being encrypted or not. As we will see in the next section, this is the case of the JWE approach.

19.5.2 A Joint Watermarking-Encryption (JWE) Approach

Security of medical images, as of any medical data, is expressed in terms of confidentiality, reliability and availability (see Sect. 19.2). Medical data reliability is important for medical practice, as it can be seen as one component of information quality establishing a degree of trust in the data practitioners' handle. The Joint Watermarking-Encryption (JWE) approach, originally studied by Bouslimi et al. in the framework of medical imaging [12, 13], has been proposed for the purpose of ensuring confidentiality (by encryption) and reliability control (by watermarking) of data, these ones being encrypted or not. The basic idea is to merge watermarking and encryption into a single (joint) operation in order to give access to a message, some image security attributes (e.g., reliability proofs and so on), available in both domains. But, as we will see, this is not the only benefit of this approach.

With JWE, one can gain in time computation since it is not necessary to decrypt the image to access the watermark, a costly operation in particular when huge volumes of data are considered. At the same time, being able to access image security attributes while not decrypting the image, maintains confidentiality protection continuity. Another benefit of merging watermarking and encryption into a single algorithm is that it is not necessary to know the decryption key when accessing the watermark into both spatial and encrypted domains. So, image reliability can be verified by entities that are not authorized to access the image content, but allowed to verify its reliability. Such situations can be found in image sharing frameworks with multiple network intermediaries (e.g., information systems or others network nodes), or within an information system where one can then check if electronic patient records are really those of the patient, without accessing their content.

One last benefit of joint watermarking-encryption compared to other approaches based on invariant or partial encryption (see Sect. 19.5.1.3) is that the host signal is fully encrypted. However, such a solution has also its constraints. Once an image is jointly watermarked-encrypted, the watermark available in the encrypted domain is read-only and cannot be modified.

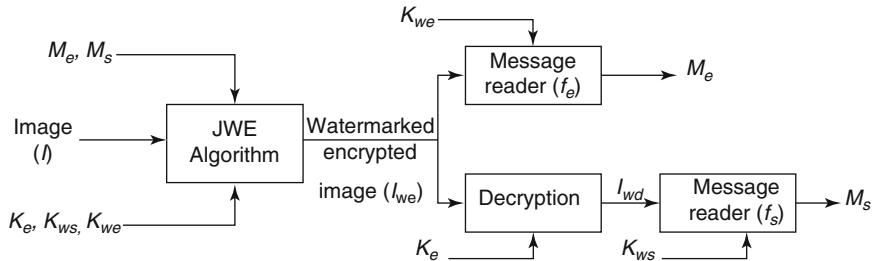


Fig. 19.9 General system architecture of a JWE algorithm. M_e , M_s , K_e , K_{ws} and K_{we} are the message available in the encrypted domain, the message available in the spatial domain, the encryption and the watermarking keys in the spatial and the encrypted domain, respectively

In this section, we detail a joint watermarking-encryption approach which merges Quantization Index Modulation (QIM) with an encryption algorithm: a stream cipher algorithm (e.g., RC4) or a block cipher algorithm (e.g., AES). Notice that using AES in CBC mode makes this approach compliant with the DICOM standard.

19.5.2.1 General Principles of the JWE System

As depicted in Fig. 19.9, a JWE algorithm conducts data encryption and watermarking in a single operation process and allows the user to insert two messages, M_s and M_e , that will be available in the spatial and encrypted domains, respectively. If encryption ensures data confidentiality, watermarking on its side is used so as to verify data integrity and authenticity.

To sum up, if we consider the JWE function W_{emb} , the joint watermarked-encrypted version I_{we} of an image I is given as:

$$I_{we} = W_{emb}(I, M_s, M_e, K_e, K_{ws}, K_{we})$$

where K_e , K_{ws} and K_{we} are the encryption and the watermarking keys in the spatial and encrypted domains, respectively. Two watermarking functions f_e and f_s are considered so as to extract embedded messages M_e and M_s from I_{we} and its decrypted version I_{wd} , respectively: $M_e = f_e(I_{we}, K_{we})$; $M_s = f_s(I_{wd}, K_{ws})$.

19.5.2.2 JWE System for Verifying Image Reliability

In practice, M_s and M_e contain some security attributes that are not necessarily the same. Nevertheless, in the case one wants to assess the image reliability in each domain, these attributes must correspond to an authenticity code AC , which identifies the image origin (e.g., about 600 bits by combining the French National Identifier with the DICOM Unique Identifier [12]), and an integrity proof. In the

spatial domain, integrity is ensured by making use of a secure hash function (e.g., SHA) computed on the image bit subset that is not modified by the watermarking process. Consider nmb this subset of bits, the message available in the spatial domain, M_s , is thus defined as: $M_s = \langle AC, SHA(nmb) \rangle$.

In the encrypted domain, integrity can be controlled by verifying the presence of a secret pseudo random sequence of bits generated using a secret watermarking key. One alternative proposed in [12] consists in considering the integrity of the watermarked-encrypted image is valid if one can retrieve the bits of such a secret random sequence at some specific locations within the SHA signature of each watermarked-encrypted block bytes. Working with the SHA, rather than directly on the encrypted bit-stream, increases the detection performance of the system in case of data modification (the reader should refer to [12] for more details). In consequence, the verification of the image authenticity and integrity in the encrypted domain, relies on extracting M_e given by: $M_e = \langle AC, PRNG(K_w^e) \rangle$.

19.5.2.3 JWE Implementation Based on QIM

The following implementation merges Quantization Index modulation (QIM) and an encryption algorithm E , which can be a stream cipher algorithm (e.g., RC4) or a block cipher algorithm (e.g., AES). It allows inserting one message m_{ei} of M_e and one message m_{si} of M_s within a block X_i of the image I .

A. QIM Adapted to Encryption

In order to insert simultaneously m_{ei} and m_{si} within X_i , and to avoid any interferences between them, Bouslimi et al. propose to adapt the QIM according to the following principles [12, 13]. Each codebook $C_{m_{si}}$ is decomposed into sub-codebooks $C_{m_{si}m_{ej}}$ such as:

$$C_{m_{si}} = \bigcup_{j=1}^q C_{m_{si}m_{ej}}$$

$$C_{m_{si}m_{ej}} \cap C_{m_{si}m_{ek}} = \emptyset, j \neq k$$

where m_{ej} is a message issued from a finite set of possible messages $M_2 = \{m_{ej}\}_{j=1,\dots,q}$. Thus, m_{si} and m_{ej} are embedded simultaneously within X_i by replacing it with X_{iw} , which corresponds to the nearest element of X in $C_{m_{si}m_{ej}}$. The sub-codebook construction is intimately linked to the encryption algorithm. Considering an encryption algorithm E and its encryption key K_e , sub-codebooks $C_{m_{si}m_{ej}}$ are built so as to verify:

$$C_{m_{si}m_{ej}} = \{Y \in C_{m_{si}} / f_e(E(Y, K_e), K_{we}) = m_{ej}\} \quad (19.6)$$

The choice of the function f_e is closely related to the goals to be achieved by the JWE system in the encrypted domain. To assess image reliability in the encrypted domain, according to the strategy depicted in [12], f_e is given as:

$$f_e(E(Y, K_e), K_{we}) = h_k$$

where h_k corresponds to the k th bit of H , the hash of the encrypted version of Y , i.e., $E(Y, K_e)$. The choice of the rank k depends on the watermarking key in the encrypted domain K_w^e .

B. Implementation with a Cipher Algorithm: The AES in CBC Mode and the RC4

Depending on the selected encryption algorithm, some other constraints have to be considered when building the subcodebooks $C_{m_{si}m_{ej}}$. In the case E corresponds to the AES in CBC mode (see Sect. 19.3.2.1), Y^e is given by [see Eq. (19.6)]: $Y^e = AES(Y \oplus X_{-1}^e, K_e)$, where X_{-1}^e is the previously encrypted block of bytes or set of pixels. So, the construction of $C_{m_{si}m_{ej}}$ depends also on the previous encrypted block. Unlike the AES, the RC4 encrypts each byte separately (see Sect. 19.3.2.2). When it is used, Y^e is given by [see Eq. (19.6)]: $Y^e = y_1^e, \dots, y_i^e, \dots, y_n^e$ with $y_i^e = y_i \oplus k_i$, where k_i corresponds to the i th byte of the keystream k generated by the RC4 according to the secret key K_e . Thus, the decomposition of each codebook into sub-codebooks depends also on the keystream bytes which are different.

Notice that the size N of each block X_i depends on the image bit depth and of the adopted encryption algorithm. Indeed, in the case of an 16 bit encoded image, because the AES works with blocks of 16, 24 or 32 bytes, N will be equal to 8, 12 and 16 pixels, respectively.

19.5.2.4 JWE Approach Performance and DICOM Interoperability

This approach is compliant with the DICOM standard in the case it works with the symmetric cipher algorithm AES in CBC mode of operation and the cryptographic hash function SHA, both recommended by this standard. To evaluate its performance, this approach has been tested over: 100 ultrasound images of 576×690 pixels of 8 bit depth, and 200 positron emission tomography (PET) images of 144×144 pixels of 16 bit depth. An example of the obtained results is provided in Fig. 19.10. The capacity rate depends on the size N of the block X_i . Since in this experiment one bit of each message is embedded within a block, capacity rates achieved in each domain are of $1/N$ bpp. The corresponding capacities are about 24,000 and 2592 bits for ultrasound and PET images, when N equals 16 and 8, respectively. They are large enough compared to the requirements for verifying the reliability of an image, which is about 1,000 bits (one digital signature—160 bits—and one

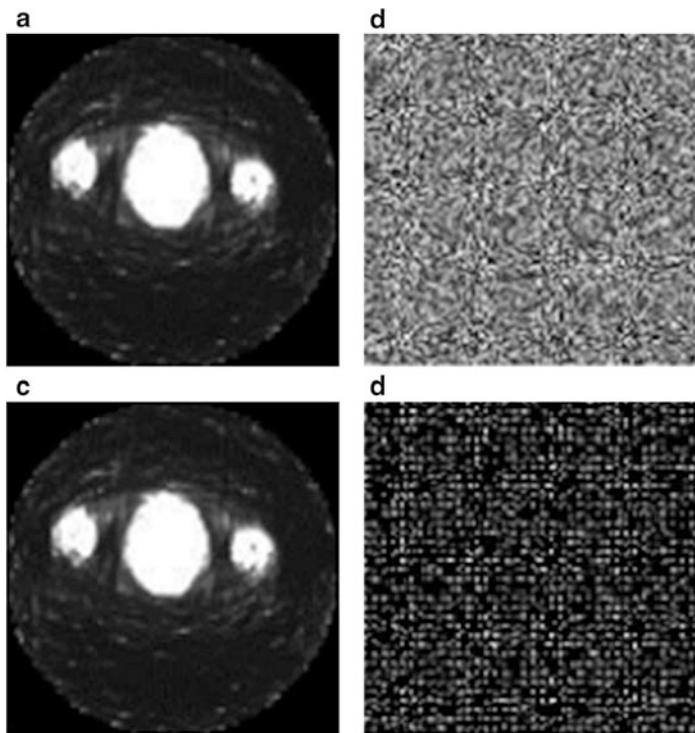


Fig. 19.10 Examples of the images used for evaluation (using AES): (a) original PET image, (b) joint watermarked/ciphered image, (c) deciphered watermark image, and (d) difference between the original image and the decrypted watermark image

authenticity code—600 bits). With these capacities and working with the AES in CBC mode, obtained PSNR values are about 60 dB and 105.26 dB for ultrasound and PET images. The image distortion is thus low.

Nevertheless, the joint watermarking/encryption process is on average two times longer than the “original” encryption algorithm (i.e., the RC4 or the AES in CBC mode). This difference is caused by the sub-codebook construction. However, it is important to notice that it offers access to new security functionalities (i.e., verifying the integrity of encrypted data without accessing their original content and so on) and that the execution time for image decryption is not impacted: it is the same as for simply encrypted images (i.e., not watermarked). For more details, the reader must refer to [12].

19.6 Conclusion

In this chapter, we gave an overview of the security needs for medical images during their storage, transmission, sharing. In particular, when looking at digital content, healthcare professionals must ensure data confidentiality, availability and reliability. We also put in evidence the complementarity of encryption and watermarking mechanisms, and the interest to merge them in order to provide an *a priori* and *a posteriori* protection at the same time, watermarking being used so as to offer security services (e.g., reliability control) even if the data are encrypted. Different approaches have been discussed but their use and adaptation to the medical domain is at its very beginning, and they have to be further explored. We also presented a very recent joint watermarking/encryption approach that we think can find real practical use in healthcare. By giving access to two distinct messages in the spatial domain and the encrypted domains, respectively, one can verify the image reliability even though it is encrypted. Due to the use of the AES in CBC mode, this approach is also compliant with the DICOM standard, in the sense that even if a system has no watermarking functionalities it can still handle the images.

Today, one of the upcoming great challenges comes with the emergence of ‘big data’ [40] poses a great challenge regarding their security. for which actual data protection mechanisms remain limited because of the amount of information to manage. That is also the case of encryption and watermarking and, in years to come, new solutions have to be developed.

References

1. Abbas, A., Khan, S.: A review on the state-of-the-art privacy-preserving approaches in the e-health clouds. *IEEE J. Biomed. Health Inform.* **18**(4), 1431–1441 (2014)
2. Adelsbach, A., Huber, U., Sadeghi, A.: Fingercasting-joint fingerprinting and decryption of broadcast messages. *Transactions on DHMS II. Lecture Notes in Computer Science*, vol. 4499 pp. 1–34. Springer, Heidelberg (2007)
3. Alattar, A.M.: Reversible watermark using the difference expansion of a generalized integer transform. *IEEE Trans. Image Process.* **13**(8), 1147–1156 (2004)
4. Allaërt, F.A., Dusserre, L.: Security of health information system in France: what we do will no longer be different from what we tell. *Int. J. Biomed. Comput.* **35**, 201–204 (1994)
5. Allaërt, A., Quantin, C.: Responsabilités et rémunérations des actes de télé-expertise. *Journal de Gestion et d’Economie Médicales* **30**(4), 219–229 (2012)
6. Alliance, W.F.: Wi-fi Protected Access: Strong, Standards-Based, Interoperable Security for Today’s Wi-Fi Networks, pp. 492–495. White paper, University of Cape Town (2003)
7. Anderson, R.J., Manifavas, C.: Chameleon: a new kind of stream cipher. In: 4th Int. Workshop Fast Software Encryption (FSE), pp. 107–113 (1997)
8. Antonini, M., Barlaud, M., Mathieu, P., Daubechies, I.: Image coding using wavelet transform. *IEEE Trans. Image Process.* **1**(2), 205–220 (1992)
9. Bender, W., Gruhl, D., Morimoto, N., Lu, A.: Techniques for data hiding. *IBM Syst. J.* **35**(3,4), 313–336 (1996)

10. Boho, A., Van Wallendael, G., Dooms, A., De Cock, J., Braeckman, G., Schelkens, P., Preneel, B., Van de Walle, R.: End-to-end security for video distribution, the combination of encryption, watermarking, and video adaptation. *IEEE Signal Process. Mag.* **30**(2), 97–107 (2013)
11. Bouslimi, D., Coatrieux, G., Roux, C.: A telemedicine protocol based on watermarking evidences for identification of liabilities in case of litigation. In: IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom), pp. 506–509 (2012)
12. Bouslimi, D., Coatrieux, G., Cozic, M., Roux, C.: A joint encryption/watermarking system for verifying the reliability of medical images. *IEEE Trans. Inf. Technol. Biomed.* **16**, 891–899 (2012)
13. Bouslimi, D., Coatrieux, G., Roux, C.: A joint encryption/watermarking algorithm for verifying the reliability of medical images: application to echographic images. *Comput. Methods Programs Biomed.* **106**, 47–54 (2012)
14. Cancellaro, M., Battisti, F., Carli, M., Boato, G., De Natale, F.G.B., Neri, A.: A joint digital watermarking and encryption method. *SPIE* **6819**, 68, 191C (2008)
15. Cancellaro, M., Battisti, F., Carli, M., Boato, G., De Natale, F.G.B., Neri, A.: A commutative digital image watermarking and encryption method in the tree structured haar transform domain. *Signal Process. Image Commun.* **26**(1), 1–12 (2011)
16. Celik, M., Sharma, G., Tekalp, A.M., Saber, E.: Reversible data hiding. *Proc. IEEE ICIP* **2**, 157–160 (2002)
17. Celik, M., Lemma, A., Katzenbeisser, S., van der Veen, M.: Look-up table based secure client-side embedding for spread-spectrum watermarks. *IEEE Trans. Inf. Forensics Secur.* **3**(3), 475–487 (2008)
18. Charpentier, A., Fontaine, C., Furion, T., Cox, I.: An asymmetric fingerprinting scheme based on tardos codes. In: 13th International Conference on Information Hiding (IH), vol. 6958, pp. 43–58 (2011)
19. Chakrabarti, C., Vishwanath, M., Owens, R.: Architectures for wavelet transforms: a survey. *J. VLSI Signal Process. Syst. Signal Image Video Technol.* **14**(2), 171–192 (1996)
20. Chen, B., Wornell, G.: Quantization Index modulation: a class of provably good methods for digital watermarking and information embedding. *IEEE Trans. Inf. Theory* **47**(4), 1423–1443 (2001)
21. Chen, C., Wu, X.: An angle QIM watermarking algorithm based on Watson perceptual model. In: Fourth International Conference on Image and Graphics, pp. 324–328 (2007)
22. Chiang, K., Chang-Chien, K., Chang, R., Yen, H.: Tamper detection and restoring system for medical images using wavelet-based reversible data embedding. *J. Digit. Imaging* **21**(1), 77–90 (2008)
23. Chrysochos, E., Fotopoulos, V., Skodras, A.N., Xenos, M.: Reversible image watermarking based on histogram modification. In: 11th Panhellenic Conf. of Informatics (PCI), pp. 93–104 (2007)
24. Costa, M.: Writing on dirty paper. *IEEE Trans. Inf. Theory* **58**, 782–800 (2003)
25. Coatrieux, G., Maître, H., Sankur, B., Rolland, Y., Collorec, R.: Relevance of watermarking in medical imaging. In: IEEE EMBS Conference on Information Technology Applications in Biomedicine, pp. 250–255 (2000)
26. Coatrieux, G., Maître, H., Rolland, Y.: Tatouage d’images médicales: perception d’une marque. In: SETIT (2003)
27. Coatrieux, G., Puentes, J., Lecornu, L., Cheze Le Rest, C., Roux, C.: Compliant secured specialized electronic patient record platform. In: 1st International Transdisciplinary Conference on Distributed Diagnosis and Home Healthcare (D2H2), pp. 156–159 (2006)
28. Coatrieux, G., Le Guillou, C., Cauvin, J.M., Roux, C.: Reversible watermarking for knowledge digest embedding and reliability control in medical images. *IEEE Trans. Inf. Technol. Biomed.* **13**(2), 158–165 (2009)

29. Coatrieux, G., Huang, H., Shu, H., Roux, C., Luo, L.: A watermarking based medical image integrity control system and an image moment signature for tampering characterization. *IEEE J. Biomed. Health Inform.* **17**(6), 1057–1067 (2013)
30. Cox, I., Miller, M., Bloom, J., Honsinger, C.: Digital Watermarking, vol. 53. Springer, Berlin (2002)
31. Cox, I., Kilian, J., Leighton, F., Shamoon, T.: Secure spread spectrum watermarking for multimedia. *IEEE Trans. Image Process.* **6**(12), 1673–1687 (1997)
32. Cox, I., Miller, M., Bloom, J., Fridrich, J., Kalker, T.: Digital Watermarking and Steganography. Morgan Kaufmann, Los Altos (2007)
33. Daemen, J., Rijmen, V.: AES Proposal: The Rijndael Block Cipher. Technical report, Proton World Int.1, Katholieke Universiteit Leuven, ESAT-COSIC, Belgique (2002)
34. David, S., Mendelson, M.: RSNA Image Share 2012. In: SIIM (2012)
35. De Moor, G.: Standardisation in health care informatics and telematics in Europe: CEN TC 251 activities. *Stud. Health Technol. Inform.* **6**, 1–13 (1993)
36. De Vleeschouwer, C., Delaigle, J., Macq, B.: Circular interpretation of bijective transformations in lossless watermarking for media asset management. *IEEE Trans. Multimedia* **5**(1), 97–105 (2003)
37. Décret n° 2010-1229 du 19 octobre 2010 relatif à la télémédecine
38. Devlin, B., Cote, L.: Data warehouse: from architecture to implementation. Addison-Wesley Longman Publishing Co., Amsterdam (1996)
39. Dierks, T., Rescorla, E.: The Transport Layer Security (TLS) Protocol, Version 1.2 (2008)
40. Diebold, F.: A Personal Perspective on the Origin(s) and Development of ‘Big Data’: The Phenomenon, the Term, and the Discipline. Sage, London (2012)
41. Dusserre, L. and Allaërt, F.A. and Quantin, C.: Pmsi et déontologie. Le Pmsi : questions juridiques et éthiques. In: Actes du colloque Information Médicale, Secret et droits d'accès des acteurs, pp. 75–79 (1991)
42. Eastlake III, D., Jones, P.: Us secure hash algorithm 1 (sha1). Technical report, RFC 3174 (Informational) (2001)
43. Eskicioglu, A., Fisher, P.: Image quality measures and their performance. *IEEE Trans. Commun.* **43**(12), 2959–2965 (1995)
44. Fontaine, C., Galand, F.: A survey of homomorphic encryption for nonspecialists. *EURASIP J. Inf. Secur.* **2007**(1), 1–15 (2007)
45. Franco-Contreras, J., Coatrieux, G., Cuppens, F., Roux, C., Cuppens-Boulahia, N.: Robust lossless watermarking of relational databases based on circular histogram modulation. *IEEE Trans. Inf. Forensics Secur.* **9**(3), 397–410 (2014)
46. Fridrich, J.: Steganography in Digital Media: Principles, Algorithms, and Applications. Cambridge University Press, Cambridge (2009)
47. Fridrich, J., Goljan, M.: Protection of digital images using self embedding. In: Symposium on Content Security and Data Hiding in Digital Media (1999)
48. Fridrich, J., Goljan, J., Du, R.: Invertible authentication. In: Proceedings of International Conference SPIE, Security and Watermarking of Multimedia Content, pp. 197–208 (2001)
49. Fung, C., Gortan, A., Godoy, Jr., W.: A review study on image digital watermarking. In: The Tenth International Conference on Networks (ICN 2011), pp. 24–28 (2011)
50. Furun, T., Bas, P.: Broken arrows. *EURASIP J. Inf. Secur.* **2008**, 3 (2008)
51. Guru, J., Damecha, H.: Digital watermarking classification: a survey. *Int. J. Comput. Sci. Trends Technol.* **2**, 8–13
52. Herrera-Joancomartí, J., Katzenbeisser, S., Megías, J., Mingüellón, D., Pommer, A., Steinebach, M., Uhl, A.: ECRYPT European network of excellence in cryptology. First summary report on hybrid systems, D.WVL.5 (2005)
53. Honsinger, C., Jones, P., Rabbani, M., Stoffel, J.: Lossless recovery of an original image containing embedded data. US Patent application, Docket No.: 77102/E-D (1999)

54. Hwang, H., Kim, H., Sachnev, V., Joo, S.: Reversible watermarking method using optimal histogram pair shifting based on prediction and sorting. *KSII Trans. Internet Inf. Syst.* **4**, 655 (2010)
55. Katzenbeisser, S., Petitcolas, F.: Information Hiding Techniques for Steganography and Digital Watermarking. Artech House, Boston (2000)
56. Kutter, M., Petitcolas, F.A.P.: A fair benchmark for image watermarking systems. In: Electronic Imaging '99. Security and Watermarking of Multimedia Contents, vol. 3657 (1999)
57. Lashkari, A., Towhidi, F., Hosseini, R.: Wired equivalent privacy (WEP). In: International Conference on Future Computer and Communication (ICFCC), pp. 492–495 (2009)
58. Légifrance: Loi n° 2002-303 du 4 mars 2002. <http://www.legifrance.gouv.fr/affichTexte.do?> (2002)
59. Liu, K.: Human visual system based watermarking for color images. In: Fifth International Conference on Information Assurance and Security (IAS), vol. 2, pp. 623–626. IEEE, New York (2009)
60. Loytnoja, M., Cvejic, N., Lahetkangas, E., Seppanen, T.: Audio encryption using fragile watermarking. In: Proc. IEEE Int. Conference on Information, Communications and Signal Processing, pp. 881–885 (2005)
61. Lee, H., Kim, H., Kwon, K., Lee, J.: Digital watermarking of medical image using ROI information. In: Proc. IEEE Int. Workshop Healthcom, pp. 404–407 (2005)
62. Lee, S., Yoo, C., Kalker, T.: Reversible image watermarking based on integer-to-integer wavelet transform information forensics and security. *IEEE Trans. Inf. Forensics Secur.* **2**(3), 321–330 (2007)
63. Li, Q., Cox, I.: Using perceptual models to improve fidelity and provide resistance to volumetric scaling for quantization index modulation watermarking. *IEEE Trans. Inf. Forensics Secur.* **2**(2), 127–139 (2007)
64. Liew, S.C., Zain, J.M.: Reversible tamper localization and recovery watermarking scheme with secure hash. *Eur. J. Sci. Res.* **49**(2), 249–264 (2011)
65. Manasrah, T., Al-Haj, A.: Management of medical images using wavelets-based multi-watermarking algorithm. In: International Conference on Innovations in Information Technology, pp. 697–701 (2008)
66. Meerwald, P., Uhl, A.: Survey of wavelet-domain watermarking algorithms. In: Photonics West 2001-Electronic Imaging. International Society for Optics and Photonics, pp. 505–516 (2001)
67. Memon, N., Wong, P.: A buyer-seller watermarking protocol. *IEEE Trans. Image Process.* **10**, 643–649 (2001)
68. Menezes, A.J., Van Oorshot, P., Vanstone, S.: Handbook of Applied Cryptography. CRC Press, Boca Raton (1996)
69. Mintzer, F., Lotspeich, J., Morimoto, N.: Safeguarding digital library contents and users: digital watermarking. *D-Lib Mag.* (1997)
70. Ni, Z., Shi, Y., Ansari, N., Wei, S.: Reversible data hiding. *Proc. IEEE Int. Symp. Circuits Syst.* **2**, 912–915 (2003)
71. Paillier-99, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Proceedings of the International Conference on the Theory and Application of Cryptographic Techniques (EUROCRYPT), vol. 1592, pp. 223–238 (1999)
72. Pan, W., Coatrieux, G., Montagner, J., Cuppens, N., Cuppens, F., Roux, C.: Medical Image Integrity Control Combining Digital Signature and Lossless Watermarking. Lecture Notes in Computer Science, vol. 5939, pp. 153–162. Springer, Heidelberg (2009)
73. Pan, W., Coatrieux, G., Montagner, J., Cuppens, N., Cuppens, F., Roux, C.: Reversible watermarking based on invariant image classification and dynamical error histogram shifting. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4477–4480 (2011)
74. Pérez-González, F., Mosquera, C., Barni, M., Abrardo, A.: Rational dither modulation: a high-rate data-hiding method invariant to gain attacks. *IEEE Trans. Signal Process.* **53**(10), 3960–3975 (2005)

75. Piva, A., Barni, M., Bartolini, F., De Rosa, A.: Data hiding technologies for digital radiography. *Proc. IEE Vision Image Signal Process.* **152**(5), 604–610 (2005)
76. Puech, W., Rodrigues, J.: A new crypto-watermarking method for medical images safe transfer. In: Proc. of the 12th European Signal Processing Conference (EUSIPCO'04), pp. 1481–1484 (2004)
77. Puech, W., Chaumont, M., Strauss, O.: A reversible data hiding method for encrypted images. In: SPIE Electronic Imaging, Security, Forensics, Steganography, and Watermarking of Multimedia Contents, p. 68191E (2008)
78. Quantin, C., Jaquet-Chiffelle, D., Coatrieux, G., Benzenine, E., Allaert, F.: Medical record search engines, using pseudonymised patient identity: an alternative to centralised medical records. *Int. J. Med. Inform.* **80**(2), e6–11 (2011)
79. Rajendra-Acharya, U., Acharya, D., Subbanna Bhat, P., Niranjan, U.: Compact storage of medical images with patient information. *IEEE Trans. Inf. Technol. Biomed.* **5**(4), 320–323 (2001)
80. Rajendra-Acharya, U., Niranjan, U., Iyengar, S., Kannathal, N., Min, L.: Simultaneous storage of patient information with medical images in the frequency domain. *Comput. Methods Programs Biomed.* **76**, 13–19 (2004)
81. Rao, K., Yip, P.: Discrete Cosine Transform: Algorithms, Advantages, Applications. Academic, San Diego (1990)
82. Rial, A., Deng, M., Bianchi, T., Piva, A., Preneel, B.: A provably secure anonymous buyer-seller watermarking protocol. *IEEE Trans. Inf. Forensics Secur.* **5**(4), 920–931 (2010)
83. Rodrigues, J., Puech, W., Fiorio, C.: Lossless crypto-data hiding in medical images without increasing the original image size. In: 2nd International Conference on Advances in Medical Signal and Information Processing, pp. 358–365 (2004)
84. Sachnev, V., Kim, H., Nam, J., Suresh, S., Shi, Y.: Reversible watermarking algorithm using sorting and prediction. *IEEE Trans. Circuit Syst. Video Technol.* **19**(7), 989–999 (2009)
85. Schmitz, R., Li, S., Grecos, C., Zhang, X.: A new approach to commutative watermarking encryption. In: 13th Joint IFIP TC6 and TC11 Conf. Communications and Multimedia Security (2012)
86. Schneier, B.: Applied Cryptography: Protocols, Algorithms, and Source Code in C. International Thomson Publishing, Paris (1997)
87. Shih, F.: Digital Watermarking and Steganography: Fundamentals and Techniques. CRC Press, Boca Raton (2012)
88. Singh, J., Dubey, A.: MPEG-2 video watermarking using quantization index modulation. In: IEEE 4th International Conference on Internet Multimedia Services Architecture and Application(IMSAA), pp. 1–6 (2010)
89. Singhal, N., Lee, Y., Kim, C., Lee, S.: Robust image watermarking based on local Zernike moments. In: IEEE 9th Workshop on Multimedia Signal Processing, pp. 401–404 (2007)
90. Smith, S.W.: The discrete fourier transform, Chap. 8. In: The Scientist and Engineer's Guide to Digital Signal Processing, 2nd edn. California Technical Publishing, San Diego (1999)
91. Snyder, A., Weaver, A.: The e-logistics of securing distributed medical data. In: Proceedings of IEEE International Conference on Industrial Informatics (INDIN 2003), pp. 207–216. IEEE, New York (2003)
92. Subramanyam, A., Emmanuel, S., Kankanhalli, M.: Robust watermarking of compressed and encrypted JPEG2000 images. *IEEE Trans. Multimedia* **14**(3), 703–716 (2012)
93. Tian, J.: Reversible data embedding using a difference expansion,. *IEEE Trans. Circuits Syst. Video Technol.* **13**(8), 890–896 (2003)
94. Watson, A.B.: DCT quantization matrices optimized for individual images. In: Human Vision, Visual Processing, and Digital Display IV (SPIE'1993), vol. 1913, pp. 202–216 (1993)
95. Xuan, G., Chen, J., Zhu, J., Shi, Y., Ni, Z., Su, W.: Lossless data hiding based on integer wavelet transform. In: Proc. MMSP, pp. 312–315 (2002)

96. Xuan, G., Yao, Q., Yang, C., Gao, J.: Lossless data hiding using histogram shifting method based on integer wavelets. IWDW 2006. Lecture Notes in Computer Science, vol. 4283, pp. 323–332. Springer, Heidelberg (2006)
97. Zhang, X.: Separable reversible data hiding in encrypted image. IEEE Trans. Inf. Forensics Secur. 7(2), 826–832 (2012)
98. Zhu, X.: A new spatial perceptual mask for image watermarking. In: 19th International Conference on Pattern Recognition (ICPR), pp. 1–4. IEEE, New York (2008)

Chapter 20

Privacy Considerations and Techniques for Neuroimages

Nakeisha Schimke and John Hale

Abstract This chapter presents key issues and techniques for preserving privacy of neuroimage datasets. Neuroimages pose unique challenges to medical privacy and image datasets. In addition to PII commonly found in medical image meta data, the data in a neuroimage itself may contain distinctive facial features that permit the identification of a subject. The chapter begins with a discussion of the identifying properties of neuroimage data, along with established processes and popular tools for storage, analysis and processing. Next, privacy risks and hazards of neuroimages are explored, including those posed by meta data and data alike. Volume rendering and facial reconstruction provide platforms for re-identification of sanitized neuroimage data. Tools for scrubbing meta data and for defacing neuroimage data are essential components to neuroimage archive and repository workflow pipelines.

20.1 Introduction

As the demand for neuroimage repositories and databases grows, so does the importance of protecting subjects and ensuring their privacy. Large scale interorganizational research efforts have the potential to transform neuroscience by fostering an open scientific environment and making datasets available for peer review and further study. However, the task of protecting subject privacy and implementing HIPAA (Health Insurance Portability and Accountability Act) [24] compliance measures can be overwhelming. While HIPAA-related measures seek to maintain subject privacy, they can encumber the researcher with tedious tasks. This burden can be alleviated through the development and implementation of an efficient solution for removing identifiable facial features from neuroimages and through integrating the process into an existing workflow to minimize the impact on the researcher.

N. Schimke • J. Hale (✉)

Tandy School of Computer Science, The University of Tulsa, Oklahoma, Tulsa, OK 74104 USA
e-mail: nakeisha-schimke@utulsa.edu; john-hale@utulsa.edu

There are two potential exploits of facial approximations that threaten anonymity: (a) *verifying a subject's identity from a list of potential subjects* (confirmation), and (b) *formulating the subject's general appearance and creating a profile for potential subjects* (novel identification).

While these applications of forensic reconstruction may require advanced software and training, it is conceivable that structural data could be used to identify a patient, thereby violating the HIPAA Privacy Rule.

This is of particular concern when a subject is a member of a small subset of a population, as is often the case in neuroimaging studies. Though re-identification may seem unlikely, a publicly available “sanitized” dataset was recently used to dispute a diagnosis given in the case of a rare disorder. A young woman claimed to be diagnosed with a rare disorder following a flu shot. She announced her case to the media, which subsequently ran a story disclosing geographic location and a general timeline. The scientific community doubted the diagnosis, and with these details, searched the Vaccine Adverse Event Reporting System (VAERS)—a publicly accessible, but sanitized database. Her particular entry in the VAERS database was found. Given the rarity of the symptoms, time of diagnosis, age and metropolitan area, it was an easy exercise to locate her entry, including evaluations that questioned the claimed diagnosis, instead ascribing it to psychogenic origins [30, 34, 44].

This improbable case highlights the need for re-evaluation of the guidelines for protecting patient data and what can be used to identify the patient. Current measures, including HIPAA, are vague on what exactly should be considered protected health information (PHI). Names, dates, and unique identifiers are established PHI, but in the aforementioned case, other data were aggregated to identify the individual. The subject's approximate age, state of residence, and an estimated timeline helped narrow down the list of potential entries, already shortened to those mentioning the rare disorder. This case demonstrates the need for stronger de-identification requirements for shared datasets in order to maintain the highest possible level of privacy.

Redacting the PHI metadata may not sufficiently de-identify the data: the neuroimage may also identify the individual. High resolution images contain detailed facial features that can be used to create an approximation of the subject's likeness. To maintain subject anonymity and meet HIPAA requirements, neuroimages must be defaced prior to sharing. Existing defacing efforts simply skull strip the neuroimages [56], which has the potential to remove brain tissue and often requires manual tuning of parameters. Other approaches are computationally intensive but leave the entire brain volume intact [10].

This chapter presents key issues and techniques for preserving privacy of neuroimage datasets. Identifying and vital properties of neuroimage data are described, along with processes and tools for their storage, analysis and processing. Privacy risks are posed by exploitation of neuroimage data and meta data alike. Any comprehensive privacy preserving solution for neuroimage datasets should consider both to satisfy legal and regulatory requirements in research and medicine.

20.2 Neuroimage Data

As a primary modality for neuroimage acquisition, structural magnetic resonance imaging (MRI) has led to numerous advancements in medicine and an improved understanding of both structure and function of human physiology and neurology [35, 60]. However, the infrastructure to acquire and use MR images is expensive. It relies on costly equipment and requires specialized resources to operate. Modern scanners produce high resolution images that demand large amounts of storage space [56, 57]. Once the data is stored, analysis can be computationally expensive.

In addition to the infrastructure related challenges, there are other limitations. It is often difficult to find subjects, and small studies may not have the statistical power that increased enrollment would provide. Subject diversity may also impact statistical power [14, 57]. Subjects must be compared once a set of images are obtained, and there will be slight variations in the scans between subjects and sessions.

There are several widely used software packages for processing anatomical images [3, 12, 19, 20]. These tools provide core functionality for the images themselves (viewing, skull-stripping, alignment, 3D rendering, landmark identification, spatial registration, surface mapping, transformation to common space, region of interest identification, etc) and several are extensible. FreeSurfer [19] from Harvard is a tool for the analysis of cortical and sub-cortical structures. BrainVoyager [12] is a commercial tool for analysis and visualization of both structural and functional neuroimages. The freely available AFNI [3] suite is primarily for the processing, analysis and display of functional MRI (fMRI) data to map brain activity rather than structural anatomy. FSL [20] from Oxford is also freely available, and comprises a library of tools for fMRI, MRI, and DTI analysis and display.

Another challenge in neuroimage analysis is the slight anatomical differences between individual brains that can interfere with results. One method for resolving these differences is to transform the brain volume to a standard coordinate system [26]. By mapping well known anatomical landmarks from an individual subject's brain to a standard brain, neuroscientists can evaluate neuroimages between subjects more easily.

The sharing and reuse of datasets can reduce some of the upfront costs involved. Collaboration enables institutions to pool data to increase the number of scans and diversity of subjects, increasing statistical power and reliability. Large scale collaboration and inter-organizational access to datasets will increase the flow of information, reducing the upfront costs of a large neuroimaging effort and meeting the requirements for publicly available data set forth by the NIH and neuroimaging journals [32]. Data sharing also enables external validation of neuroimaging studies.

The discussion concerning privacy in neuroimaging data has focused on two distinct entities: (1) the researcher and the scientific community, and (2) the patient [53]. Data sharing raises concerns over intellectual property, but it also requires quality control and release of study data to prevent the misuse of data. Patient privacy requires the removal of personally identifiable information so that

the subject cannot be tied to the study. The community is aware of the privacy issues, but there is still a need for further examination of the privacy concerns in neuroimages. While some journals and data repositories require neuroimages to be sanitized before submission, there is no standard for neuroimage de-identification.

Also pressing are the rules set forth by HIPAA/HITECH requiring strict measures to manage protected health information (PHI). These specify items considered to be PHI, and while most of these items are well-defined (e.g. names, dates, social security numbers), they also include “full face photographic images and any comparable images” and any other unique identifying number, characteristic or code [24].

The specification for the DICOM (Digital Imaging and Communications in Medicine) standard allows manufacturers to specify tags to be automatically embedded within a DICOM image [31]. These tags range from scanner and acquisition parameters to the subject’s personal data, including names, identifiers, and dates. It is rather straightforward to remove identifiable well-defined fields, but it is a more challenging task to determine and redact vague identifiers. However, the relevance of the original neuroimage may not always be retained when redacting data; for some studies, it may degrade the quality of the data to remove particular details, such as age, that are relevant to the study.

A larger question looms: Are neuroimages themselves considered PHI? And if so, how can they be shared without breaching the privacy of the subject? Decidedly, neuroimage data can hold visual features that can be used to uniquely identify a subject. Structural MRIs, for example, are of such a resolution that high fidelity facial reconstruction is well within the realm of possibility [43]. Approaches to preserving privacy for medical images include strategic removal or transformation of identifying features.

20.3 Privacy Risks with Medical Images

This section explores the privacy risks in neuroimage datasets and how these datasets can be exploited to violate individual privacy and/or to re-identify a subject. It discusses the current state of facial reconstruction and recognition and their accuracy with respect to re-identification.

20.3.1 Neuroimage Privacy Threat Scenarios

Neuroimages contain a wealth of information about a subject. The meta data may contain personally identifying information (PII). The neuroimage data itself may contain unique features capable of identifying a subject. In addition, the data may reveal conditions, diagnoses or features deemed private information. In

DICOM Image Database (DICOM tag)			
Occupation (0010,2180)	Birth Date (0010,0090)	Gender (0010,0040)	Address (zip code) (0010,1040)
Nurse	1959	M	75287
welder	1966	F	75335
Entrepreneur	1966	F	75074
Firefighter	1955	M	75261
Auditor	1978	M	75261

Patient Name	Occupation	Birth Date	Gender	Zip Code	Diagnosis
Katie	Entrepreneur	04/18/1966	F	75074	Sprained ankle
Andrew		01/06/1953	M	75038	Concussion
Alice		08/20/1978	F	75240	Bronchitis
Ben		05/04/1982	M	75038	Asthma

Fig. 20.1 Re-linkage using an imaging database

isolation, knowledge of these attributes may not constitute a privacy breach, but in combination they commonly do.

There are four potential conditions of neuroimages (and medical images in general) that can be combined to breach subject privacy:

1. The image reveals a condition or diagnosis.
2. The image is used to link the patient to other records.
3. The existence of the image is used to infer a condition or diagnosis.
4. The image contains uniquely identifying features.

The first is the straightforward case where an image directly reveals a condition, such as an x-ray revealing a broken bone or a neuroimage revealing a tumor. This, combined with the ability to associate a neuroimage with a patient identity, constitutes a breach of privacy. The second case addresses the re-linkage problem, where multiple “de-identified” datasets are aggregated to re-identify a subject. In one exercise, “anonymized” medical records from the Massachusetts Group Insurance Commission were linked to voter registration lists [51]. Using the overlapping fields (birth date, zip code, sex), the medical records of the Massachusetts governor were identified. An example of re-linkage in a medical imaging context is shown in Fig. 20.1.

An existential threat may occur without ever viewing the actual image, presuming the image can be connected to an individual. For example, a subject who is part of a neuroimaging study may be assumed to have a condition even though they are part of a control group. Similarly, it may be inferred that a patient who receives a diagnostic scan has an injury.

In the fourth case, the image itself contains inherently identifiable information. If a uniquely identifying landmark or anomaly is present, it may be difficult to remove without sacrificing the usefulness of the image. This also includes facial features present in high resolution neuroimages, which can be effectively redacted without damaging the image. In this case, removing identifying meta data is not a solution—the neuroimage data itself poses the problem.

20.3.2 Volume Rendering and Facial Recognition

In high resolution structural neuroimages, facial features may be highly visible, allowing for a volume rendering of the entire head that results in a face. These volume renderings can be exploited in two ways: (1) the approximation can be used to create a profile of the unknown subject to narrow a potential list of possible subjects, or (2) the subject's appearance can be used to identify him by visual inspection or through facial recognition. Either of these cases relies heavily on the ability to create a realistic and accurate approximation of the subject. This is a non-trivial task, particularly when computerized forensic reconstruction methods attempt to automate a manual task that is often artistic in nature and therefore subjective. Facial recognition, too, is hindered by algorithms that do not perform as well as human recognition. However, the ability to create an image that is instantly recognizable as a human face requires further consideration.

The ability to recreate medical images in three-dimensions is a powerful analysis tool. Imaging software provides the ability to interact with the resulting volumes. When applied to structural neuroimages, the result is a face, shown in Fig. 20.2.

The canonical medical image package, 3D Slicer [1], provides a wide range of parameters. Threshold levels, color, and opacity can be specified, and various lighting situations can be simulated. Other packages, such as the AFNI Render Volume Plugin [33] and MRIcron [39], offer similar functionality.

Facial reconstruction is a forensic technique used to reconstruct a likeness of the subject from the skull. Rather than exact reproduction, its purpose is to produce an approximation that can be used to narrow down a list of subjects or generalize his appearance [61]. Manual facial reconstructions were originally conducted in the absence of recognizable tissue, requiring estimation of tissue depths at reference

Fig. 20.2 Volume rendering from 3D Slicer



points in the face. This is as much an artistic process as it is scientific, and reconstructions may vary between artists. The use of modern imaging techniques like MRI and CT include some tissue in addition to the skull, guiding depth selection. Accompanying metadata, such as age, gender, and weight, can also be used to guide the reconstruction process to deliver a more accurate approximation.

Facial recognition can be applied using a variety of techniques to achieve either novel identification, attempting to discover an identity (or list of identities) without prior knowledge of the subject, or identity confirmation, where the subject is compared to another potential subject (or list of subjects). Metadata can be used to guide a facial recognition search, narrowing down the potential subjects using the basic non-PHI fields such as gender and age.

A fundamental disadvantage of facial recognition is the reliance on a large database of faces for comparison. It is also difficult to validate matches; the returned result may be a closer match to the subject than the actual subject. Beyond computational and storage requirements of executing facial recognition searches, collecting and maintaining such a database raises privacy concerns. The database would also need to be periodically updated and re-evaluated for maximum effectiveness.

Literature surveys [2, 11, 63] indicate that the reported performance of facial recognition is high, particularly when factors such as light, pose, or expression were tightly controlled. Performance, however, tends to decrease dramatically when increasing the size and variation of images within the database.

The current limitations and relatively poor performance of existing facial recognition techniques make it tempting to dismiss the potential for re-identification based on flawed assumptions: facial recognition will never improve, and only correct identifications are potentially problematic. The latter argument fails to consider the damage caused by an incorrect identification. While a correct identification may reveal sensitive information, challenging a false identity may require the individual to reveal their medical records.

The problems plaguing facial recognition techniques are not easily confronted nor are they impossible to solve, and researchers in the field are making progress. Facial recognition techniques are detailed with links to recent advances at the Face Recognition Homepage [22]. A NIST report on face recognition illustrates the improvements in the field with a significant decrease in the false non-match rate and false match rates from 1993 to 2010 [23].

Hardware advances can also improve the state of facial recognition. Increased storage capacity and computing power allow higher resolution images to be stored and compared more quickly. Facial recognition software tends to struggle when viewing angles and lighting vary [63]. Volume rendering software for structural neuroimages can generate multiple images with a wide range of light sources and angles to match the source photograph.

Novel solutions, such as CAPTCHA,¹ used to combat similar issues with optical character recognition (OCR) are being adapted for use in facial recognition. This

¹The CAPTCHA system presents a test that is difficult for a machine but easy for humans [4, 21].

idea has been applied to images as the Google Image Labeler [5], using human brain power in a game to label images, an area with which computer vision techniques struggle. Facebook offers social authentication [38], a CAPTCHA that requires users to correctly identify their friends in pictures. These approaches harness the capabilities of human vision and recognition to bridge the performance of computer vision techniques, blending automated and manual approaches. However, these solutions rely on correct image tagging and therefore may be prone to errors introduced by tagging.

20.3.3 Re-identification Using Structural MRI

An ideal case for re-identification would occur when an MRI series is obtained for a specific patient, and volume rendering is performed on the dataset. The resulting face is used to search a large database of faces, and a match is found using one of any number of facial recognition techniques. This resulting match is validated by visual inspection.

There are a number of obstacles before this use case becomes a typical scenario. The following sections describe this use case and what challenges are faced at each step, offering alternative examples that circumvent these challenges, demonstrating the need for implementing privacy controls rather than waiting until the “ideal” use case can be easily exploited.

Software packages like 3D Slicer [1] offer basic volume rendering and require little previous experience with facial reconstruction or volume rendering. While these approximations may not be accurate, they can be enough to ascribe an identity to a subject. The impact of incorrect re-identifications should also be considered. If reconstruction and recognition techniques are perceived as sound and proven technologies, then the resulting approximations and matches will be given more weight, even if they are inexact or incorrect.

There are three primary categories of facial recognition algorithms [46, 63]: (1) holistic approaches using the whole face, (2) feature-based approaches using specific features (e.g., eyes, nose, mouth, etc.), and (3) hybrid approaches that combine both holistic and feature-based techniques. In any of these three cases, removing the specific features (as opposed to removing the entire face region) used for feature-based techniques, should significantly degrade performance of facial recognition as well as reconstruction.

MRI-based approaches could help mitigate some of the problems with current facial recognition systems. For example, various lighting and angles could be simulated from the reconstruction to match the target photographs. Advancements in computational power and storage space make solutions such as this a distinct possibility to improve upon current resource-limited approaches.

Concerns about the feasibility of facial recognition applied to neuroimages are largely the same as the issues with photograph-based recognition. The language of the HIPAA Privacy Rule classifies “full facial photographic or comparable

images” as PHI. Thus, if neuroimage-based recognition can elicit similar results to photographs, they must be protected as if they were facial photographs by HIPAA bound entities.

Neuroimage datasets typically contain valuable metadata. Even if the accompanying metadata is not considered PHI, it can still be used to reveal the identity of the subject. Data such as gender, ancestry, and age are not immediately apparent from the MRI, nor are they particularly identifying on their own in most cases; however, when combined with the image data, they can be applied to assist in finding or validating a match.

It has been demonstrated that seemingly innocuous data can be used to re-identify individuals when applied to small subsets of the population [37, 44]. In the VAERS case, the metadata alone was enough to find a probable match, and though there was never any confirmation as to the subject’s identity, it was enough to damage the claims of the subject. (This raises another interesting scenario for MRI re-identification: if facial recognition and reconstruction techniques are viewed by the public as technologically sound, then incorrect re-identification can be equally harmful to a suspected subject as an accurate re-identification.)

20.4 Privacy Preservation Techniques for Medical Images

This section describes techniques and approaches to preserving privacy in medical images, and neuroimage datasets specifically. It discusses standard de-identification techniques, including skull stripping and defacing processes, relates them to prevailing regulatory standards for patient privacy, and offers metrics and characteristics of sound privacy preserving tools for neuroimages. This section also presents a view of popular neuroimage archival and collaboration initiatives, and the privacy considerations and solutions associated therewith.

20.4.1 De-Identification Techniques

While facial reconstruction techniques are not foolproof, it is reasonable to assume that with external knowledge such as diagnoses or geographic region of research institution, a reconstruction could be used to narrow the list of potential patients or to confirm a patient’s identity. With this in mind, it is necessary to implement a technique that maintains the highest possible level of privacy for patients without encumbering the researcher. An automated defacing effort would benefit both research institutions and clinical practices, ensuring that sanitization is always performed but without interfering with the task of collecting or analyzing the images.

Several techniques have been proposed to de-identify structural MRI data, but none have yet emerged as a standard method. These techniques range from removing

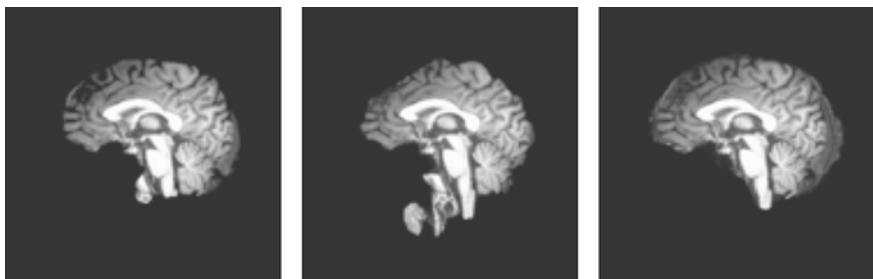


Fig. 20.3 Skull Stripping: 3dSkullstrip in AFNI (*left*); BET in FSL (*middle*); HWA in Freesurfer (*right*) [41]

all non-brain tissue (skull stripping) to removing only facial features (defacing). Skull stripping is the process of segmenting brain and non-brain elements to remove extraneous tissue (such as eyes, skin, etc) and bone that may interfere with analyses [18, 45, 50]. Skull stripping techniques rely on the ability to differentiate between brain and non-brain tissue and are used for improved registration, cortical segmentation, and cortical flattening. They are particularly useful for structural MRI due to the high spatial resolution of acquired images. The primary challenge with any of the skull stripping methods is the difficulty of brain segmentation. It is often a subjective process even when performed manually, and manual segmentation can take up to two hours to perform [18, 50]. Using a manual approach for correcting an automated segmentation may offer a balance between efficiency and accuracy, but is still costly in terms of time and resources. Figure 20.3 shows the results of skull stripping a structural MRI image using three tools.

In contrast to skull-stripping, defacing is a de-identification process in which the goal is to remove only facial features. Defacing techniques should remove a sufficient amount of non-brain tissue to render the subject's face unidentifiable and unable to be reconstructed. The main objective is de-identification rather than segmentation, and these algorithms can be conservative by including anything that may be brain tissue and only discarding areas that have zero probability of being brain. This avoids some of the more complex regions, such as the optical nerves that can often complicate the process [50]. However, skull stripping techniques developed for a whole brain volume may be sensitive to removal of non-brain tissue and defaced volumes may be more challenging to skull strip. Skull stripping parameters may also be chosen to favor the specific region of the study at the expense of other regions. There are also potentially desirable features, such as CSF and electrode placement, that may be removed by skull stripping [10]. Thus, it is preferable to retain as much non-identifying tissue as possible to maximize data reusability.

The MRI Defacer in [10] defaces images by removing identifying features only, leaving the brain and most of the surrounding non-brain tissue intact. The algorithm relies on an atlas constructed by manual labeling of facial features. An optimal linear transform is computed for the input volume. It then creates a mask of all voxels with

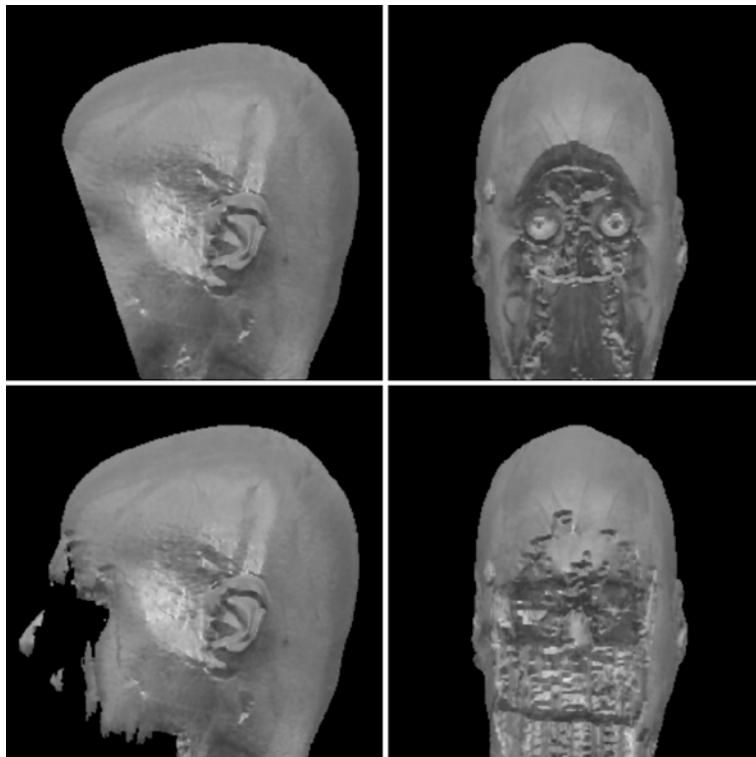


Fig. 20.4 Defacing: Quickshear (*top*); MRI Defacer (*bottom*) [41]

a non-zero probability of being brain tissue, preserving voxels and surrounding areas with any possibility of being brain tissue. The remaining voxels are then removed if they have a non-zero probability of being a facial feature, thus removing only facial features with zero probability of being brain tissue. In general, this solution does not negatively impact subsequent skull stripping techniques, and, in some cases, improves it. This approach relies on an atlas constructed using T1-weighted data and therefore is only valid for T1-weighted images. Other images, such as T2-weighted images, require a specific face atlas for each type. The results of this method are shown at the bottom of Fig. 20.4 (sagittal—left; coronal—right).

Other defacing approaches seek more efficient strategies for de-identifying neuroimage data. The Quickshear Defacing method uses a convex hull to identify a plane that divides the volume into two parts, one containing facial features and another the brain volume, and removes the voxels on the facial features side [42]. This method, the results of which are shown at the top in Fig. 20.4 (sagittal—left; coronal—right) is a practical alternative to competing solutions and can provide performance gains and preserve valuable data.

Adoption of any of these de-identification techniques must be predicated on an objective evaluation and consideration of relevant regulatory and compliance requirements. The HIPAA Privacy Rule provides a mechanism for research purposes through which some PHI can be preserved as a “limited dataset” [24], containing PHI relevant to the study but not enough to be identifiable. Limited datasets may contain geographical information (city, state, zip), dates and other identifying numbers, characteristics, or codes, but must exclude facial photographs. HIPAA deems a dataset to be sufficiently de-identified in either of the following circumstances:

1. A formal determination by a qualified statistician.
2. The removal of specified identifiers of the individual and of the individual’s relatives, household members, and employers is required, and is adequate only if the covered entity has no actual knowledge that the remaining information could be used to identify the individual [24].

The Privacy Rule does not quantify the acceptable limits of identifiability, though it states that zip codes must be excluded if they contain fewer than 20,000 people.

Determining an acceptable threshold is a difficult task, especially as computational power and storage capacities improve. The HIPAA threshold of 20,000 people per zip code is considerably narrowed when accounting for other metadata. Applying a filter for gender would roughly halve that number to 10,000 [55], and other subsets of the population may be even more vulnerable to metadata filtering. Defining a threshold that considers minimal exposure for variably sized populations may be more difficult than simply omitting or restricting the metadata. It can be said that facial approximations generally do not produce an exact likeness of the original subject. However, the resulting facial reconstruction can yield enough information to create a general profile for the subject which can then be used for identification.

The primary objective of defacing is feature removal to obscure the subject’s identity and sufficiently reduce the risk of re-identification. This can be done by removing all non-brain tissue or only tissue that can be used in facial reconstruction and recognition. *Feature removal* refers to the removal of a whole feature (e.g. the nose), and *partial feature removal* refers to the removal of a portion of a single feature (e.g. the tip of the nose).

It is tempting to assume that, because skull stripping is already integrated into the analysis workflow, no further de-identification is necessary. However, skull stripping algorithms are widely varied, and different approaches or parameter sets can yield different volumes. Inconsistencies or inaccurate parameter selection may lead to loss of desired brain tissue, as well as difficulty when performing meta-analysis and other future studies that rely on uniform preprocessing.

Some defacing techniques attempt to remove all non-brain tissue with a non-zero probability of being facial tissue. Like the limited dataset, it may not be necessary to completely redact all identifiable features when removing a subset of the features can render the subject unrecognizable. Removing more features may not always be desirable, but the inherent risk of leaving features can be evaluated based on the purpose of data sharing. The goal of defacing is to render the subject unrecognizable to both human perception and automated facial recognition software. It is not

Table 20.1 Landmarks used in facial reconstruction [13, 15, 36, 52] and corresponding feature memberships

	Reference number	Landmark	Feature
1	1	Supraglabella	External
2	2	Glabella	External
3	3	Nasion	Nose
4	4	End of nasals	Nose
5	5	Mid-philtrum	Mouth
6	6	Upper lip margin	Mouth
7	7	Lower lip margin	Mouth
8	8	Chin-lip fold	External
9	9	Mental eminence	External
10	10	Beneath chin	External
11	32/11	Frontal eminence	External
12	33/12	Supraorbital	Eyes
13	36/15	Suborbital	Eyes
14	37/16	Inferior malar	Eyes
15	46/25	Lateral orbit	Eyes
16	45/24	Zygomatic arch, midway	External
17	44/23	Supraglenoid	Mouth
18	51/30	Gonion	External
19	47/26	Supra M^2	Mouth
20	49/28	Occlusal line	Mouth
21	50/29	Sub M_2	Mouth
–	34/13	Lateral glabella	Eye
–	35/14	Lateral nasal	Nose
–	38/17	Lateral nostril	Nose
–	39/18	Naso-labial ridge	Mouth
–	40/19	Supra canina	Mouth
–	41/20	Sub canina	Mouth
–	42/21	Mental tubercle ant.	External
–	43/22	Mid lateral orbit	Eyes
–	48/27	Mid-masseter muscle	External
–	52/31	Mid mandibular angle	External

The left column refers to [36], and the second column to [15]. The far right column indicates feature classification

necessary to remove all facial features to obscure the subject's identity, but enough features should be removed so that the remaining features are not enough to re-identify the individual or reconstruct the removed features.

When considering the three categories of automated facial recognition techniques (whole face, feature based, and hybrid), feature removal can significantly impact the performance of all three approaches. The resulting reconstruction should also be unrecognizable upon visual inspection. Facial features should not be able to be recreated using symmetry or any residual tissue.

There are a number of facial landmarks that are used in facial reconstruction [36, 52] and facial recognition [46]. In the former, facial landmarks represent anatomical points on the skull that can be reliably located and where soft tissue depths are known. Commonly used cranial landmarks (32 including both left and right points according to Phillips and Smut, and 52 according to De Greef et al.) with well-researched tissue depths are listed in Table 20.1. The additional landmarks appearing in De Greef et al. are based on reliability of location and inclusion in previous studies [15]. The goal of defacing is to obscure a subset of these landmarks so that the remaining calculable ratios violate uniqueness between subjects.

Studies in facial recognition by humans can provide guidance in determining which features to remove. One study suggests that faces are recognized by individual features, particularly the eyes, nose, mouth, and eyebrows, though recognition is also influenced by holistic processing [49]. These features are grouped into two primary categories: internal features (eyes, nose, mouth) and external features (hair and jawline), which are traditionally considered less reliable [47, 48]. It has been shown that configuration, or the relative placement of the internal features, plays an important role in human face recognition [9, 48].

Another study demonstrated that subjects match familiar faces on internal features more quickly than unfamiliar faces, but matching on external features shows no difference between familiar and unfamiliar faces [62]. This agreed with the findings of an earlier study by Ellis et al. that found both internal and external features important in identifying unfamiliar faces, with internal features more significant than external for familiar faces [17].

A robust defacing solution should validate the results to ensure that (1) only non-brain tissue is removed and all brain tissue is preserved, and (2) that the remaining facial tissue is not sufficient to create an approximation or compare using facial recognition. It should be noted that the latter does not ensure that re-identification will not be possible. In some cases, the PHI in the limited dataset may be enough to identify the individual even without the MRI.

With freely available software, it is easy to create a facial image from an MRI, shown in Fig. 20.2, even without previous neuroimaging experience. While the image is instantly recognizable as a human face, the subject's identity may not be easily discerned. However, an inexact rendering may lead to incorrect re-identification that can potentially be damaging to the subject and the “re-identified” individual.

For neuroimaging studies, the number of subjects enrolled is often small [40, 58, 59] and frequently includes small subsets of the population, such as those with a specific disorder or from a specific age group. It is necessary to carefully evaluate the limited dataset to ensure that re-identification is not possible from the metadata, even without the addition of the image data. Even if the metadata is enough to re-identify the individual, an unredacted neuroimage can be used to lend credibility to the re-identification.

The following traits are desirable in a defacing technique but are not considered primary characteristics:

- **Legacy processing.** Defacing techniques should work on legacy datasets as well as those produced by modern scanners. Skull stripping mechanisms may not perform as well on legacy data due to the wide variation in quality [18].
- **Flexibility and transparency.** Flexibility and openness in methods used to identify brain tissue and remove non-brain tissue create an agile environment that can easily adapt to future improvements in both re-identification and de-identification.
- **Ease of use.** An effective de-identification method should run without user intervention and only require interaction when an error is encountered. The implementation should detect potential hazards, such as brain tissue removal or impartial feature removal, and notify the user.
- **Data management.** The creation of multiple subsets of data may introduce inconsistencies in the datasets. It may also increase the potential of a privacy breach if the researcher inadvertently shares the original dataset rather than the de-identified version. It is therefore preferable to have a data management component that interfaces with the defacing algorithm to ensure that only the defaced dataset is shared with external collaborators.

A performance metric for a general neuroimage de-identification technique focuses on the two primary objectives of defacing: Facial Feature Removal (FFR) and Brain Volume Preservation (BVP). There are many more factors to crafting successful de-identification approach as discussed in the previous section, but these are the most significant factors and can be easily quantified.

There are two potential methods for determining a defacing score: a binary score that indicates whether or not a face is present, and a more granular score that reflects the percentage of voxels that belong to facial features. The latter is a more specific quantification, but it is difficult to implement. It requires a labeled face atlas and a transformation to atlas space. As with the MRI Defacer face atlas, different acquisition methods may require different atlases, and a transform to a canonical space may not be completely accurate. The measure may also be skewed by the accuracy of the atlas. The atlas would need to be hand labeled, and this is a fairly subjective process that may not abstract well to other datasets. Tailoring an atlas to a specific dataset may be more time consuming than manually checking images to see if the facial features are sufficiently removed. A voxel-based percentage may benefit from appropriate weighting of various voxels belonging to a particular feature.

A binary score sacrifices granularity for speed and versatility. This method is easier to implement and does not rely on an atlas. It uses face detection mechanisms to determine if a face is present in the resulting volume rendering. Manual user inspection can be used instead of an automated approach. This does not offer the specificity of a voxel-based quantification, and the face detection methods may not be completely accurate. False negatives may be more complicated to detect in a binary score than a voxel-based score.

The binary scoring method does not require a resource intensive transform or rely on a previously hand-labeled atlas or weighting, and therefore it is a more general solution. It uses the Viola-Jones detector implemented in OpenCV and the corresponding training set. The facial feature removal score is

$$S_F = \begin{cases} 0 & \text{if a face is detected} \\ 1 & \text{otherwise} \end{cases} \quad (20.1)$$

with an ideal value of zero.

The brain volume is represented by \mathbf{B} , the set of all voxels that are classified as brain tissue. V is the number of voxels in \mathbf{B} removed by the defacing technique. The brain tissue preservation score S_B represents the percentage of voxels mistakenly discarded, with an ideal value of 0:

$$S_B = \frac{V}{B} \quad (20.2)$$

This percentage-based approach weighs all voxels that are potentially brain tissue as equally significant, including cerebrospinal fluid and other non-brain tissues identified as brain voxels by the masking technique. Moreover, a large number of voxels in the brain mask can skew the results, a problem exacerbated by the increasing trend in neuroimage resolution. For example, if the brain mask contains over one million voxels, a loss of 100,000 voxels results in a retention score of 90 %, which may seem like an adequate rating. However, a loss of 10 % of the brain volume may be unacceptable depending on the situation, and it may be more beneficial to examine the raw number of discarded voxels. This score applies only to the identifiability of the neuroimage itself as inherently identifiable through facial features and does not consider the presence of any metadata. Improper de-identification of metadata may undermine any defacing effort, and the comprehensive dataset should be evaluated for potential re-linkage.

Since distinguishing brain and non-brain tissues is a challenging problem, this metric is only as accurate as the brain mask that it is measured against. If the brain mask does not contain all desirable brain tissue, then the resulting metric may return an inflated score; conversely, a mask that contains too much extraneous tissue will deflate the score. These scores offer a metric for evaluating defacing mechanisms. For any dataset, it may be preferred to manually evaluate both the efficacy of defacing and preservation of brain tissue, but the scores can be used as a preliminary sweep to identify potentially problematic images or as a sanity check.

20.4.2 Privacy in Neuroimage Archives and Collaboration Initiatives

Initiatives and archives designed to spur large-scale research and collaboration through the sharing of neuroimage datasets must integrate privacy and security controls into their frameworks. Many establish workflows and pipelines decorated with redaction and sanitization tools for archivists and researchers.

The Alzheimer’s Disease Neuroimaging Initiative (ADNI) [6] is a successful example of multisite neuroimaging collaboration. The funding of ADNI has enabled the collection of MRI and PET images, as well as other data. By 2010, data had been collected from patients at over forty sites and distributed to over 1,300 investigators [54]. As of 2015, ADNI data has been cited in 487 publications and downloaded over 8 million times [7, 8] ADNI has produced promising results, inspiring collaborative efforts for other diseases [25].

The Laboratory of NeuroImaging (LONI) at UCLA [27] provides the infrastructure for ADNI. A web based application is used to upload data to the archive, and the data are de-identified and validated at the site-level so that sensitive patient data is not transmitted. LONI advises that face images may be reconstructed from certain MR images under HIPAA and therefore should be considered PHI. LONI also provides the LONI De-identification Debablet [28], a Java tool for removing metadata from image files with pseudonymous identifiers. For datasets that do not require project approval, by accepting the terms of use, the downloader agrees that they will not attempt to re-identify the data.

The fMRI Data Center is a repository for peer-reviewed neuroimages and related study data published in the *Journal of Cognitive Neuroscience* [56]. While no longer open for new data submissions, the subject privacy policy holds researchers responsible for sanitizing their own data prior to submission, but the Data Center also performs a scan of submissions to remove any potentially identifying data overlooked by the researcher. Researchers may also submit skull stripped images to prevent reconstruction of the subject’s face; otherwise, the Data Center will apply skull stripping to the images before they are made available. This dual approach adequately protects the privacy of the subject but the researcher must either perform sanitization or let the Data Center bear the responsibility. It has the potential to strip data it determines to be uniquely identifying without notifying the researcher and thus threatens the integrity of the data. Altering the data while it is not under control of the researcher may have wider implications on the quality of the dataset.

The Extensible Neuroimaging Archive Toolkit (XNAT) developed at Washington University in St. Louis [29] is an open software framework that provides data management and control for neuroimaging studies and related data. XNAT is designed to facilitate multi-institutional research collaboration through data management and control. Data flows through XNAT in five phases: (1) Data Acquisition, (2) Quarantine, (3) Local Use, (4) Collaboration, and (5) Public Access. In the Data Acquisition phase, data is uploaded to XNAT and is subsequently quarantined. There, the user validates it before viewing and analyzing it. Users can then create subsets of the data that can be sent to collaborators or released for public viewing.

The current version of XNAT provides basic support for sanitizing patient data. The DICOMBrowser tool [16] allows users to view or edit DICOM metadata both manually and with batch processing. It is offered in both a graphical and command-line interface. The flexibility of XNAT allows researchers to upload data in multiple formats with varying study parameters. However, this presents a challenge when standardizing a sanitization process; it requires the end user to tailor the anonymization process to a particular study or session through scripts based on the DICOM tag key-value pairs.

20.5 Conclusion

Neuroimages and structural MRI are playing an increasingly prominent role in diagnostic medicine. Moreover, collaborative research initiatives can transform neuroscience by fostering an open scientific environment that makes neuroimage datasets available for peer review and study.

The collection, use and disclosure of neuroimage data encumbers serious consequences for individual privacy. Neuroimage datasets constitute PHI in two ways: (i) the meta data associated with them contains personal and revealing demographic information and (ii) the data itself may contain identifying features. In addition, neuroimage datasets may reveal medical conditions or other characteristics an individual may wish to remain confidential. In some cases, the mere knowledge of the very existence of such data for an individual might constitute a violation of privacy.

Frameworks and platforms for storing and sharing neuroimage data must, accordingly, integrate privacy preserving solutions geared toward image data. Such technologies either sanitize the personally identifying meta data associated with neuroimages or transform the neuroimage data itself so that re-identification is difficult or impossible. Their selection and adoption should be based on objective performance metrics that are meaningful in the context of prevailing regulatory guidelines and compliance standards.

References

1. 3D Slicer. <http://www.slicer.org/> (2010)
2. Abate, A.F., Nappi, M., Riccio, D., Sabatino, G.: 2D and 3D face recognition: a survey. *Pattern Recogn. Lett.* **28**, 1885–1906 (2007)
3. AFNI. <http://afni.nimh.nih.gov/> (2009)
4. von Ahn, L., Blum, M., Hopper, N.J., Langford, J.: CAPTCHA: using hard AI problems for security. In: *Proceedings of Eurocrypt* (2003)
5. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2004)

6. Alzheimer's Disease Neuroimaging Initiative: ADNI: Alzheimer's disease neuroimaging initiative. <http://www.loni.ucla.edu/ADNI/> (2010)
7. Alzheimer's disease neuroimaging initiative: ADNI data publications. <http://adni.loni.usc.edu/news-publications/publications/> (2015)
8. Alzheimer's disease neuroimaging initiative: ADNI data usage stats. <http://adni.loni.usc.edu/data-samples/adni-data-usage-stats/> (2015)
9. Balas, B.J., Sinha, P.: Portraits and perception: configural information in creating and recognizing face images. *Spat. Vis.* **21**(1–2), 119–135 (2007)
10. Bischoff-Grethe, A., Ozyurt, I.B., Busa, E., Quinn, B.T., Fennema-Notestine, C., Clark, C.P., Morris, S., Bondi, M.W., Jernigan, T.L., Dale, A.M., Brown, G.G., Fischl, B.: A technique for the deidentification of structural brain MR images. *Hum. Brain Mapp.* **28**, 892–903 (2007)
11. Bowyer, K.W., Chang, K., Flynn, P.: A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. *Comput. Vis. Image Underst.* **101**(1), 1–15 (2006)
12. BrainVoyager. <http://www.brainvoyager.com/> (2010)
13. Claes, P., Vandermeulen, D., De Greef, S., Willems, G., Clement, J.G., Suetens, P.: Computerized craniofacial reconstruction: conceptual framework and review. *Forensic Sci. Int.* **201**(1–3), 138–145 (2010)
14. Costafreda, S.G.: Pooling fMRI data: meta-analysis, mega-analysis and multi-center studies. *Front. Neuroinformatics* **3**, 33 (2009)
15. De Greef, S., Claes, P., Vandermeulen, D., Mollemans, W., Suetens, P., Willems, G.: Large-scale in-vivo Caucasian facial soft tissue thickness database for craniofacial reconstruction. *Forensic Sci. Int.* **159S**, S126–S146 (2006)
16. DicomBrowser. <http://nrg.wustl.edu/projects/DICOM/DicomBrowser.jsp> (2010)
17. Ellis, H.D., Shepherd, J.W., Davies, G.M.: Identification of familiar and unfamiliar faces from internal and external features: some implications for theories of face recognition. *Perception* **8**, 431–439 (1979)
18. Fennema-Notestine, C., Ozyurt, I.B., Clark, C.P., Morris, S., Bischoff-Grethe, A., Bondi, M.W., Shattuck, D.W., Leahy, R.M., Rex, D.E., Toga, A.W., Zou, K.H., Morphometry BIRN, Brown, G.G.: Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: effects of diagnosis, bias correction, and slice location. *Hum. Brain Mapp.* **27**, 99–113 (2006)
19. Free Surfer. <http://surfer.nmr.mgh.harvard.edu/> (2009)
20. FMRIB Software Library. <http://www.fmrib.ox.ac.uk/fsl/> (Accessed 2010)
21. Google: reCAPTCHA. <http://www.google.com/recaptcha> (Accessed 2011)
22. Grgic, M., Delac, K.: Face recognition homepage. <http://face-rec.org> (Accessed 2011)
23. Grother, P.J., Quinn, G.W., Phillips, P.J.: Report on the evaluation of 2D still-image face recognition algorithms. Technical report, National Institute of Standards and Technology (2010)
24. HIPAA administrative simplification: regulation text (2009)
25. Kolata, G.: Rare sharing of data leads to progress on Alzheimer's. *New York Times*, New York (2010)
26. Lancaster, J.L., Woldorff, M.G., Parsons, L.M., Liotti, M., Freitas, C.S., Rainey, L., Kochunov, P.V., Nickerson, D., Mikiten, S.A., Fox, P.T.: Automated Talairach Atlas labels for functional brain mapping. *Hum. Brain Mapp.* **10**, 120–131 (2000)
27. LONI: Laboratory of Neuro Imaging. <http://www.loni.ucla.edu/> (2010)
28. LONI De-identification Debablet. <http://www.loni.ucla.edu/Software/DiD> (2009)
29. Marcus, D.S., Olsen, T.R., Ramaratnam, M., Buckner, R.L.: The eXtensible neuroimaging archive toolkit: an informatics platform for managing, exploring, and sharing neuroinformatics data. *Neuroinformatics* **5**(1), 11–33 (2007)
30. Najera, R.: Records show case of dystonia is psychogenic and not related to flu vaccine. <http://www.examiner.com/x-13791-Baltimore-Disease-Prevention-Examiner> (2009)
31. National Electrical Manufacturer's Association: digital imaging and communications in medicine 2009. <ftp://medical.nema.org/medical/dicom/2009/> (2009)
32. NIH: NIH data sharing information. http://grants.nih.gov/grants/policy/data_sharing/ (2009)

33. NIMH: AFNI: using the volume rendering plugin. http://afni.nimh.nih.gov/pub/dist/edu/latest/afni_handouts/afni09_render.pdf (2009)
34. Novella, S.: Neurologica blog—the dystonia flu shot case. <http://www.theness.com/neurologicablog/?p=1152> (2009)
35. Ogawa, S., Lee, T.M., Kay, A.R., Tank, D.W.: Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc. Natl. Acad. Sci.* **87**(24), 9868–9872 (1990)
36. Phillips, V.M., Smuts, N.A.: Facial reconstruction: utilization of computerized tomography to measure facial tissue thickness in a mixed racial population. *Forensic Sci. Int.* **83**, 51–59 (1996)
37. Rescue, G.: Desiree jennings flu shot injury. http://generationrescue.org/desiree_jennings.html (2009)
38. Rice, A.: The facebook blog: a continued commitment to security. <http://www.facebook.com/blog.php?post=486790652130> (2011)
39. Rorden, C.: MRIcron. <http://www.cabiatl.com/mricron/mricron/index.html> (2010)
40. Salimi-Khorshidi, G., Smith, S.M., Keltner, J.R., Wager, T.D., Nichols, T.E.: Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. *NeuroImage* **45**, 810–823 (2009)
41. Schimke, N.: Quickshear defacing for neuroimages: algorithm, implementation and analysis. The University of Tulsa, Tulsa (2011)
42. Schimke, N., Hale, J.: Quickshear defacing for neuroimages. In: Proceedings of the 2Nd USENIX Conference on Health Security and Privacy, HealthSec'11, pp. 11–11. USENIX Association, Berkeley, CA, USA. <http://dl.acm.org/citation.cfm?id=2028026.2028037> (2011)
43. Schimke, N., Kuehler, M., Hale, J.: Preserving privacy in structural neuroimages. In: Submitted to Proceedings of the Conference on Data and Applications Security and Privacy (DBSec) (2011)
44. Science Blogs—Respectful Insolence: Has Desiree Jennings' VAERS report been found? http://scienceblogs.com/insolence/2009/11/has_desiree_jennings_vaers_report_been_f.php (2009)
45. Ségonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B.: A hybrid approach to the skull stripping problem in MRI. *NeuroImage* **22**, 1060–1075 (2004)
46. Shi, J., Samal, A., Marx, D.: How effective are landmarks and their geometry for face recognition? *Comput. Vis. Image Underst.* **102**, 117–133 (2006)
47. Sinha, P.: Identifying perceptually significant features for recognizing faces. In: Proceedings of the SPIE Electronic Imaging Symposium (2002)
48. Sinha, P., Balas, B.J., Ostrovsky, Y., Russell, R.: Face recognition by humans, chap. 8. Elsevier Academic Press, Cambridge (2006)
49. Sinha, P., Balas, B.J., Ostrovsky, Y., Russell, R.: Face recognition by humans: 19 results all computer vision researchers should know about. *Proc. IEEE* **94**(11), 1948–1962 (2006)
50. Smith, S.M.: Fast robust automated brain extraction. *Hum. Brain Mapp.* **17**, 143–155 (2002)
51. Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Uncertainty Fuzziness Knowledge Based Syst.* **10**(5), 557–570 (2002)
52. Taylor, K.T.: Forensic Art and Illustration. CRC Press, Boca Raton (2000)
53. Toga, A.W.: Neuroimage databases: the good, the bad and the ugly. *Nat. Neurosci.* **3**, 302–309 (2002)
54. Toga, A.W., Crawford, K.L., the Alzheimer's Disease Neuroimaging Initiative: The informatics core of the Alzheimer's disease neuroimaging initiative. *Alzheimers Dement.* **6**(3), 247–256 (2010)
55. US Census Bureau QuickFacts. <http://quickfacts.census.gov> (2010)
56. Van Horn, J.D., Grethe, J.S., Kostelec, P., Woordward, J.B., Aslam, J.A., Rus, D., Rockmore, D., Gazzaniga, M.S.: The functional magnetic resonance imaging data center (fMRI-DC): the challenges and rewards of large-scale databasing of neuroimaging studies. *Philos. Trans. Biol. Sci.* **356**, 1323–1339 (2001)
57. Van Horn, J.D., Toga, A.W.: Is it time to re-prioritize neuroimaging databases and digital repositories? *NeuroImage* **47**, 1720–1734 (2009)

58. Wager, T.D., Lindquist, M.A., Nichols, T.E., Kober, H., Van Snelleberg, J.X.: Evaluating the consistence and specificity of neuroimaging data using meta-analysis. *Neuroimage* **45**, S210–S221 (2009)
59. Wang, J., Conder, J.A., Blitzer, D.N., Shinkareva, S.V.: Neural representation of abstract and concrete concepts: a meta-analysis of neuroimaging studies. *Hum. Brain Mapp.* **31**, 1459–1468 (2010)
60. Westbrook, C.: *Handbook of MRI Technique*. Wiley, Blackwell (2014)
61. Wilkinson, C.: Computerized forensic facial reconstruction: a review of current systems. *Forensic Sci. Med. Pathol.* **1**(3), 173–177 (2005)
62. Young, A.W., Hay, D.C., McWeeny, K.H., Flude, B.M., Ellis, A.W.: Matching familiar and unfamiliar faces on internal and external features. *Perception* **14**, 737–746 (1985)
63. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: a literature survey. *ACM Comput. Surv.* **35**, 399–458 (2003)

Chapter 21

Data Privacy Issues with RFID in Healthcare

Peter J. Hawrylak and John Hale

Abstract Radio frequency identification (RFID) provides a means to implement the “last-mile” connection in a connected world, often referred to as the Internet of Things (IoT). RFID has been widely used in the retail and construction sectors for supply-chain management, and has provided significant benefits to those sectors. RFID has also been employed by healthcare to improve supply-chain management and monitor the locations of patients and providers to improve service offerings. The wireless and low-cost aspects of RFID introduce privacy concerns. However, there are some privacy issues related to the use of RFID, including tracking and the ability of an attacker to obtain sensitive data that is stored in the RFID tag. This chapter will explore the use of RFID in healthcare and identify issues relating to privacy that need to be addressed in these use-cases. An overview of RFID technology is presented followed by an overview of applications of RFID in healthcare. The privacy issues are then identified and potential solutions described. The privacy issues identified can be addressed using proven and standard security practices, many of which are already implemented by healthcare providers. A discussion of how to extend these practices to include RFID technology is provided.

21.1 Introduction

Radio frequency identification (RFID) is a key component of the connected world and the Internet of Things (IoT). There are many applications of RFID in medicine, from managing inventory to monitoring patient’s health. The wireless nature of RFID coupled with the requirement for a low unit price of RFID tags introduces several privacy concerns. This chapter provides a brief overview of the use-cases of RFID in medicine and then provides a thorough analysis of the privacy issues and risks associated with these systems. Solutions to these risks are hypothesized and the impact of the proposed solution is estimated on the original use-case.

P.J. Hawrylak (✉) • J. Hale

Tandy School of Computer Science, The University of Tulsa, Tulsa, Ok, USA
e-mail: peter-hawrylak@utulsa.edu; john-hale@utulsa.edu

21.1.1 *RFID as a Technology*

RFID is a form of wireless technology, which has wide adoption in supply chain management (inventory tracking), and will play a key role in the development and deployment of the Internet of Things (IoT). RFID provides the *last-mile* connection between the asset or object and the larger information system, i.e., the Internet. RFID systems consist of two basic components, the RFID tag, which is attached to assets, and the RFID reader, which provides the gateway between the tags and the larger information system (Internet).

RFID tags contain a unique identifier (UID) that is used to identify a single RFID tag out of a group of RFID tags. This is in contrast to a barcode system where the barcode represents the serial number of the item, but is not unique; the quantity of items is determined by the manual process of scanning the barcode on each item. The UID of the RFID tag can also double as an address, or IP address, for the tag and item it is associated with. In this manner RFID provides a way to assign IP addresses to objects to create the IoT. RFID tags that provide only their UID are referred to as “license plate” RFID tags as they just provide a license plate for an item but no additional information.

A RFID tag can provide higher functionality beyond the license plate tag, by including additional re-writable memory or sensors in the tag. These tags are sometimes referred to as “data-rich tags.” The additional memory can be used to store information such as product expiration dates or to provide a record of maintenance activities on an item.

RFID tags generally fall into one of three categories, passive, battery assisted passive (BAP), and active, based on how they are powered. Passive tags do not have an on-board power source and must harvest their operating energy from their ambient surroundings. Often, they harvest energy from the radio frequency (RF) signal from the RFID reader. Passive tags communicate using backscatter communication, which is low-power and does not require the tag to contain a transmitter. Tag cost and operating energy requirements are the major limiting factors in functionality provided by passive tags.

BAP tags are positioned between passive and active RFID tags in terms of functionality and cost. BAP tags include a on-board power source (e.g., a battery) for non-communication operations, such as sensing between reads. Communication is accomplished using backscatter. The inclusion of the power source allows the BAP tag to use all of the RF energy from the reader transmission for backscatter resulting in a longer read (communication) range compared to a passive tag. BAP tags are widely used applications that require sensing or monitoring because they can take readings between reads. Passive tags are only powered during the read and that is the only time they can take a reading.

Active tags are typically the most costly, have an on-board power source, and use an active (powered) transmitter and receiver. They support long-range communication and can communicate in areas unfriendly to RF communication (e.g., significant RF noise). Communication ranges vary based on the type of

transceiver and battery in the tag. Typical ranges for active tags in RFID and RTLS applications range from tens of meters to several kilometers. Active tags are widely used to monitor conditions, such as intrusion detection in shipping containers, and transmit alerts of abnormal conditions (e.g., over a satellite link). In addition, active tags are often used to track objects and people in a real-time location system (RTLS). The active transmitter and receiver support advanced location determination algorithms to be executed to determine the precise location of the tag.

There are several RFID communication protocols that are used in RFID systems. Differences between these protocols include the frequency bands used, data formatting, data encoding, and supported commands. All protocols provide a means to obtain the UID from the RFID tag, but several include support for reading and writing small user-memories on the RFID tag. These user memories are between 100-bits to several mega-bits (Mb).

From a frequency band perspective RFID systems operate in four different bands: 125–135 kHz, 13.56 MHz, 433 MHz, and 860–960 MHz. The 125–135 kHz band supports low-frequency (LF) RFID systems. These systems are typically passive and operate in the near-field. They are used for livestock and animal tracking because to the ability of the LF signals to penetrate animal tissue. The read-range (distance between the reader and tag, when the tag can be successfully read) is about 30 cm.

The 13.56 MHz band supports high-frequency (HF) RFID systems. These systems are passive RFID systems and operate in the near-field. HF RFID is used in many access control systems (badge systems), contactless payment, public transit fare collection, and supply chain management. The 13.56 MHz band is designated as an industrial, scientific and medical (ISM) band worldwide, which means that it can be used without a license anywhere in the world. This makes HF RFID a good choice for implementing systems that need to work around the world. HF RFID provides better read ability when tags are in the presence of liquid and HF RFID has been identified as one option for using RFID to meet ePedigree mandates for pharmaceuticals. HF antennas are typically designed using a loop structure and some examples of HF tags are shown in Fig. 21.1.

The 433 MHz band supports ultra-high frequency (UHF) active RFID systems. These systems are used to track large assets such as shipping containers and have a read-range of up to several hundred meters. They can communicate with the reader in environments that are not friendly to radio frequency (RF) communication (e.g., large amounts of metal such as on a cargo ship).

The 860–960 MHz band supports UHF passive RFID systems. These systems are used in supply-chain management and inventory tracking applications. There is no frequency band in this range that is available worldwide. Thus, each country uses a different frequency band in this introduces problems in the design of the antenna for the RFID tag. Specifically, it is difficult to design an antenna that will work well over the entire 100 MHz band and still meet the low price point and small area requirements for RFID tags. Often, the size of the object being tagged limits the size of the RFID tag. The EPCglobal organization manages the Electronic Product Code (EPC) numbering system that manages the tag UIDs used by different users.



Fig. 21.1 Passive HF RFID tags

The EPC number is used as a key to lookup the item in a central database where detailed information about that item is stored. Thus, it is important that each user assigns UIDs from within a set of numbers that are assigned to them. This is similar to how IP addresses are allocated and managed. UHF antennas are typically based on a dipole instead of a loop. Examples of UHF tags are shown in Fig. 21.2.

The UHF passive RFID systems also support the inclusion of batteries to create a BAP RFID system. The typical application of a UHF BAP RFID system is to attach sensors to the RFID tag to monitor conditions between read events. Table 21.1 shows frequency band, corresponding RFID system type, the applicable ISO standards, and typical applications of these systems.

The remainder of this chapter is structured as follows. Section 21.2 presents the privacy needs for medical data and provides a high-level framework for accessing privacy concerns and to define when a breach occurs. Next, Sect. 21.3 presents applications of RFID in medicine, with a focus on applications for inventory tracking, tracking people, and device management. The use of RFID in each area are presented along with example applications. Then, Sect. 21.4 describes the risks to privacy associated with these three use-cases for RFID in medicine. Section 21.5 presents some potential solutions to these issues. Finally, Sect. 21.6 concludes this chapter by summarizing the issues and solutions that were presented.

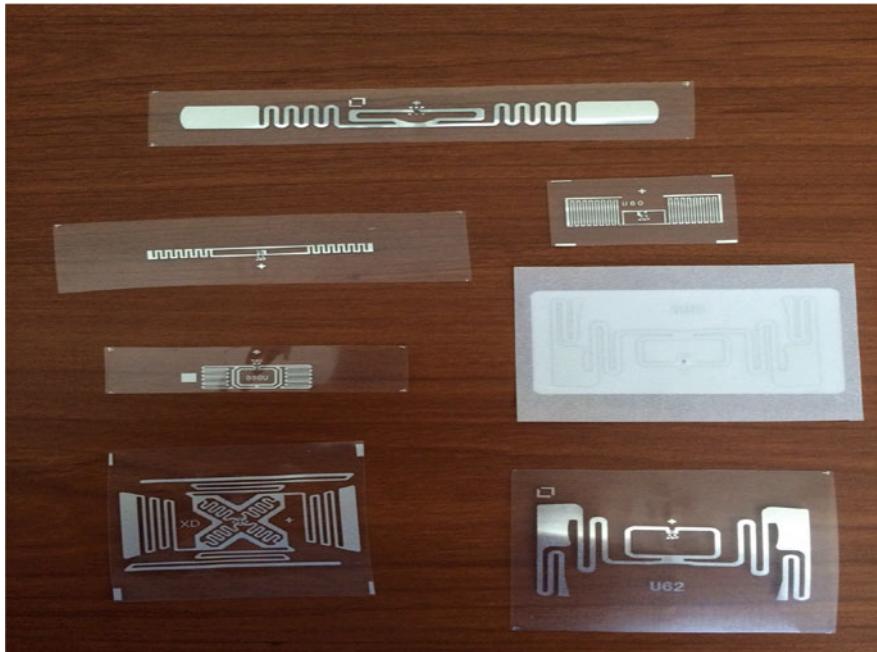


Fig. 21.2 Passive UHF RFID tags

Table 21.1 The frequency band, RFID type, corresponding ISO standard, and typical applications

Frequency band	RFID type	ISO standard	Typical application
125 kHz	LF	ISO 18000-2	<ul style="list-style-type: none"> • Livestock and pet tracking
13.5 MHz	HF	ISO 18000-3	<ul style="list-style-type: none"> • Public transit fare • Access control systems
433 MHz	UHF (active)	ISO 18000-7	<ul style="list-style-type: none"> • Tracking shipping containers
860–960 MHz	UHF (passive)	ISO 18000-6 (air interface) <ul style="list-style-type: none"> • ISO 18000-61 (Type A) • ISO 18000-62 (Type B) • ISO 18000-63 (Type C) • ISO 18000-64 (Type D) 	<ul style="list-style-type: none"> • Supply chain management

21.2 Dimensions of Privacy in Medicine

Privacy in the modern age is commonly concerned with controlling the disclosure of identifiable information. Today, medical data is often digitized and transmitted between information technology (IT) systems to facilitate diagnosis, treatment and care. Much of that data is often considered to be protected health information (PHI), as it contains sensitive and identifying information. PHI includes items such as data

of birth, home address, contact information (e.g. home address, telephone numbers, and email address), emergency contacts, treatment details, and medical history.

The foundations of medical privacy rest in principles promoted by the 1973 HEW Report, a study conducted by The Secretary's Advisory Committee on Automated Personal Data Systems within the Department of Health, Education, and Welfare [1]. The study [1] established a Code of Fair Information Practices comprising five ideals:

1. There must be no personal-data record-keeping systems whose very existence is secret.
2. There must be a way for an individual, to find out what information about him is in a record and how it is used.
3. There must be a way for an individual to prevent information about him obtained for one purpose from being used or made available for other purposes without his consent.
4. There must be a way for an individual to correct or amend a record of identifiable information about him.
5. Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take reasonable precautions to prevent misuse of the data [1].

The Code of Fair Information Practices profoundly shaped the Privacy Rule in The Health Insurance Portability and Accountability Act (HIPAA) of 1996. HIPAA was passed by the U.S. Congress with five objectives: improving portability and continuity of health insurance coverage; combating waste, healthcare fraud, and abuse; promoting medical savings accounts; improving access to long-term care services; and simplifying health insurance administration [2, 3]. HIPAA's Privacy Rule establishes conditions to preserve patient privacy in healthcare systems. Other aspects of HIPAA deal with privacy indirectly by incorporating provisions intended to preserve security properties of patient data in any form.

Concerning privacy, HIPAA prescribes how an individual's PHI can be collected, used, or disclosed. Identifying information such as billing records, data entered by healthcare providers, patient information stored in the computer system, and the health insurer's information about treatment and care, are protected by HIPAA. The Privacy Rule permits access to PHI only when providing treatment, paying medical providers for services, protecting public health, giving reports to law enforcement, or when a patient has explicitly authorized a third party. HIPAA applies to any organization that touches PHI. This includes healthcare providers, health plan providers, and business associates.

The 2009 Health Information Technology for Economic and Clinical Health (HITECH) Act gave the U.S. Department of Health and Human Services (HHS) the authority to create programs for private and secure electronic health information exchange [4]. The Omnibus Rule by the Office for Civil Rights clarified and expanded the scope of HITECH's and HIPAA's privacy and security provisions. This

rule increases the liability of business associates, broadens the right of individuals to PHI access and notice, and increases privacy protection for genetic information [5].

The Office of the National Coordinator (ONC) in HHS has adopted eight privacy principles essential for privacy solutions in healthcare information systems [6]:

1. **Individual Access:** Simple and timely access for individuals to their personal health information.
2. **Correction:** Ability of an individual to dispute and correct erroneous personal health information.
3. **Openness and Transparency:** Openness and transparency of policies, procedures and practices relating to an individual's medical records.
4. **Individual Choice:** Ability of an individual to make informed decisions regarding the collection, use and disclosure of their personal health information.
5. **Collection, Use and Disclosure Limitation:** Appropriate control of the collection, use and disclosure of personal health information governed by the necessity to accomplish a specific purpose.
6. **Data Quality and Integrity:** Reasonable effort to guarantee personal health information has not been altered inappropriately and that information is accurate.
7. **Safeguards:** Application of security controls to preserve the confidentiality, integrity and availability of medical records and personal health information.
8. **Accountability:** Monitoring and reporting of events and actions in HIEs that potentially constitute breaches, misuses or violations of privacy and security.

The wireless nature of RFID introduces and magnifies several privacy concerns. Specifically, the wireless communication link provides the attacker the ability to launch an attack from a distance. An example of such an attack is using a rogue RFID reader to retrieve (read) the data from an RFID tag that contains medical data about a patient, in order to obtain that patient's PHI. Solutions to address this concern and other potential attacks are presented later in this chapter. Further, the low-cost and low-power (especially for passive tags) requirements for the RFID tags limit the available security and privacy measures that can be implemented. However, most of these concerns can be addressed using typical techniques.

A recent analysis of security concerns for RFID tags [7] of Gen-2 and ISO 18000-63 [8] passive ultra-high frequency (UHF) tags identified four major categories of attacks: interception, interruption, modification, and fabrication. Of these interception is a direct threat to privacy because personal information could be obtained from the transmission or the intercepts used to track a person. Tracking can be used to extract further personal information about the person by linking the rooms of the medical facility they visit to the services provided by those rooms (e.g., cancer treatment). One application is to reduce waiting times [9] by keeping track of the location of patients and providers.

21.3 RFID in Medicine

RFID and RTLS offer many benefits to medicine and have been successfully applied to address several issues. A general overview of these technologies can be found in the following: for RFID [10–14], and for RTLS [15–17].

This chapter focuses on the following use-cases of RFID and RTLS technology in medicine: (a) *inventory tracking*, (b) *tracking people*, and (c) *medical device management*. These use-cases are described in further detail in the following sections.

21.3.1 Inventory Tracking

Inventory tracking and supply chain management is a major use-case for RFID, especially in the retail sector [18]. Tracking inventory and linking the use of supplies to the appropriate patient is a major issue for most hospitals. RFID has been used to create “smart cabinets” that are able to link supplies that are removed to a particular patient [11]. This enables better billing accuracy and can be linked into hospital inventory management systems to ensure that supplies are reordered when they are needed. Several users have realized savings from a reduction in the number of items that expire before use from such systems [19–21].

Another use of RFID is to monitor whether or not a patient is taking their prescribed medications. One implementation uses a specially designed drawer with a RFID reader and RFID tags on each medicine container to track when a patient removes medicine from the drawer [22]. This system records what medications were removed and replaced, but does not verify the dose of the medication taken (e.g., how many pills were removed). An expansion of this system was implemented to include the ability to record the weight of the medication container before and after removal [23]. Another system incorporates a video camera to track medication for elderly patients to ensure compliance with the prescribed instructions [24]. Another medication compliance system uses a RFID reader contained (at least the antenna) in a necklace with a tag in each pill to monitor when a pill is ingested [25]. General RFID systems integrated into hospitals can be used to help providers maintain supply levels and can potentially allow them to move closer to the just-in-time inventory process used in many retail establishments [26]. Other uses include using RFID systems to track the location and status of digital infusion pumps [27].

21.3.2 Tracking People

RFID and RTLS provide a means to track people in an environment. There are a number of means to determine the location of the RFID tag in an environment,

including range estimation based on signal strength [28–32] time-of-arrival/flight [33–36], and angle of arrival [37, 38]. Other more advanced options to get better accuracy are possible.

The medical applications in this area often focus on tracking patients of care facilities [39] and tracking locations of patients and staff members in a doctor's office for scheduling purposes. Several researchers have also investigated using RFID to locate and track patients and providers for the purpose of improving workflow planning and management [40]. The ability to track patients allows the care facility to monitor the location of at-risk patients (e.g., those suffering from memory loss diseases) and to be alerted if the patient leaves the facility or enters an a restricted or unsafe area. One study looked at the use of a commercial RTLS system to monitor the entry/exit points of a hospital [41]. Alerts are triggered when a patient passes through a entry/exit point. One benefit of monitoring only entrance and exit points is a reduction in the number of RFID readers needed to monitor a facility. This reduces the cost of the system. However, the downside of this is that the patient's location cannot be tracked inside the facility but only at entry/exit points. This is effective for many facilities because they only need to monitor the entrances and exits points and not the entire facility.

Systems have also been deployed using RFID and other Internet of Things type technologies to track how a person interacts with their environment [24, 42, 43]. These systems can be used to help manage patients suffering from cognitive diseases while allowing them to remain in their homes or have more independence compared to a traditional nursing home.

Other systems track the location of patients and can be used to help control the spread of infectious disease. One such system was implemented in a hospital in Taipei to help control the SARS virus [44] by identifying patients with SARS symptoms or diagnosed with SARS, and also by identifying the individuals that were in contact with a person newly diagnosed with SARS. This system helped to isolate those patients with SARS and those medical workers assigned to treat SARS patients from the rest of the hospital's population. Many other RTLS systems have been applied to improve processes in hospitals [9, 45–48].

21.3.3 Device Management

RTLS systems have been used to track and locate devices and equipment within a hospital [41]. These RTLS systems differ from those issuing alerts when people leave a facility without authorization in that they track the location of devices (objects) in the entire facility. One concern with wireless location systems is being able to read through walls in a building. For example, this could result in the system indicating that an asset is in room 314 when it is actually next door in room 315.

RFID provides the ability to track inventory, including medical supplies and pharmaceutical drugs. Inventory tracking using RFID is a common application of the technology in the retail sector. In the retail sector RFID is being used to reduce

the time (man-hours) required to take inventory [49, 50] and in some instances can be used to give a merchant an instant view of inventory on hand in their store. This information is then used to ensure that the optimal product selection [51] is available on the store shelves to customers and that something ordered over the Internet is actually available for pickup in the store.

Similar methods can be applied to healthcare to track medical supplies. RFID can be used to link supplies used to a particular patient, to help ensure that each patient is charged for only those supplies that they use. The hospital can use this information to improve the accuracy of their inventory and can reorder needed supplies so that they arrive just-in-time for their use. Another use of RFID is to identify the patient to verify their medications [52]. This helps reduce the occurrence of the medical issues arising from giving patient A, patient B's medications. This application also simplifies the nurse's task in delivering the medications and should result in less time being spent on delivery of the medication, allowing more time for patient interaction and providing care service.

A RFID system was pilot tested in a hospital in Cyprus to identify patients and track inventory in the pharmacy [26]. Identification of patients is a key factor in providing the correct medical treatment and is often the root-cause for medical mistakes. One such application of this technology is to link the patient to their medication to ensure that each patient receives the proper medication and dosage. In the system presented in [26] the pharmacists found benefit in the improved and quick inventory features provided by the system, which helped them to maintain the optimal supply of medication in the pharmacy.

21.4 Issues and Risks

Privacy issues differ from security in that privacy is geared to the collection, use, and release of confidential information. Security is concerned with confidentiality and also data integrity and availability. With respect to privacy, the requirement to have access to the data—availability—is often at odds with the need to maintain confidentiality of the data. For example, a secure system that provides sufficient confidentiality will provide privacy protection. However, in medicine, it is sometimes hard to enforce access control, especially in response to an emergency situation. For these emergency type situations several medical databases allow qualified medical professionals to gain access to a patient's medical information for treatment purposes, even if they do not normally have access rights to that information. Medicine introduces limits on the time required to obtain information (e.g. or risk the patient dying) that in other industries (e.g. financial institutions) are not present and allow for alternative means of authenticating the user or providing a redacted version of the data to the user to allow them to perform their task. The low cost nature of RFID tags makes the inclusion of strong security options difficult. Further, passive RFID devices must harvest enough energy to power the tag during strong, and often long, cryptographic operations. An overview of security and privacy concerns relating to RFID is found in [53].

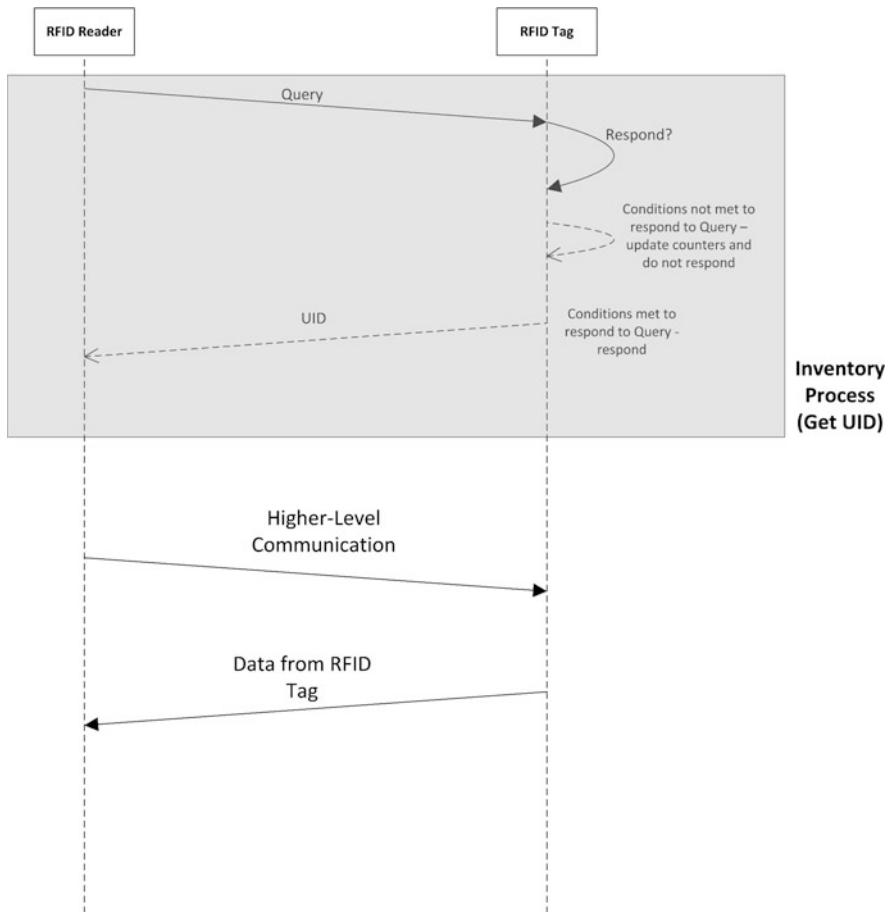


Fig. 21.3 Basic exchange between an RFID tag and reader

Tracking is difficult to prevent because the mainstream RFID protocols require each tag to have a unique identifier. A typical RFID exchange begins with the reader retrieving the UID from one tag and then using the UID to carry out a higher-level communication where data (from the user data section) of the tag is read by the reader. Alternatively, the higher-level communication session could be used to set operating or security parameters of the tag, or to write data to the tag's user memory. All tags support the capability to retrieve the UID, but not all support the higher-level functions (or have user memory). Figure 21.3 shows the basic components of the RFID exchange. The first part recovering the UID is supported by all RFID tags, while the second part (higher level communication) is supported by some tags.

There are several proposed solutions to address tracking [54–56], but most require significant infrastructure on the part of the system. This large infrastructure overhead may not be feasible for pervasive systems such as those envisioned in

the IoT. The Internet of Things concept requires every device to have a unique address (similar to an IP address) and this means that these devices will be easy to track, which is often beneficial for the Internet of Things applications, but troubling from a privacy standpoint. Ideally, the same RFID tag could be used to provide information and services from the point of manufacture (or assignment) to the point where it is discarded or returned (recycled). This is one of the goals of the EPC numbering system. New additions to the EPC Gen-2 protocol (basis for the ISO 18000-63 standard) include commands to support security features to authenticate readers to tags and to provide communication channels that ensure confidentiality and integrity of the data and higher-level commands passed between the reader and tag [57].

The EPC number provides a unique key to search in a centralized database to obtain additional information about a RFID tag and the item or person it is associated with. This process allows the RFID tag to contain only “license plate” data to link the tag to an entry in the database. Hence, RFID tags providing only a unique ID that can be used as a search key for a database are termed “license plate tags.” Figure 21.4 shows this process. Typically, a user-interface is involved to present data to the user and often times offers data entry capabilities to the user. First, the RFID tag is singulated, which means that it is identified by the reader out of the collection of RFID tags present. Once singulated, the RFID reader can retrieve the tag’s EPC number and can perform higher-level operations (e.g. read and write user-memory) with the tag. The EPC number is then used by the reader to query the central database. The central database holds detailed information about the tag and associated asset and provides this information to the user-interface. At this point, the user can request additional operations to be performed on the tag or can alter data in the database. This model separates the data storage from the RFID tag and this has benefits from a privacy perspective because strong authentication procedures can be implemented to protect the database and the tag only carries the EPC number (license plate). One drawback of this division of information is that the data may not be available to the user if they do not have a connection to the central database.

A recent survey of RFID applications in healthcare found that privacy was a significant factor in the success or failure of an RFID deployment that involves tracking individuals, but not assets [17]. The concern includes both patients and healthcare providers. While the EPC Gen-2 protocol and ISO 18000-63 standard include support to permanently disable the digital components on a RFID tag, through a kill command, using this feature prevents the RFID tag from being used in the future. Physical means can also be used to achieve similar results by damaging the antenna by including a removable piece in the antenna to reduce the read range to a few centimeters [58].

One of the difficulties is that for hospitals to adopt RFID and RTLS technology the system needs to be able to integrate with the other systems present in the hospital [59]. These systems include data entry and collection systems, central databases, and user interfaces. The data entry and collection systems need to be able to interact with the RFID readers directly or through middleware (e.g., a device

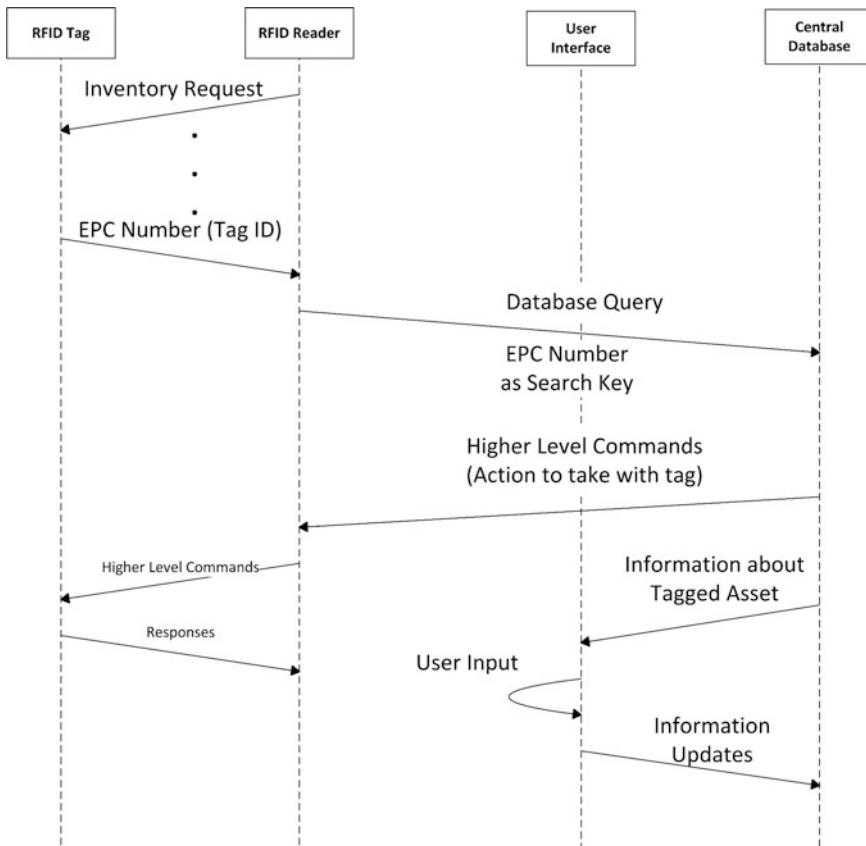


Fig. 21.4 Using the EPC number to retrieve additional information about the tag and associated asset

driver) software to have the readers establish and manage communication between the larger system and the RFID tags. The middleware or the reader can also process the RFID tag data and present it in a format expected by the hospital's larger system. The medical database needs to include additional fields to hold tag UIDs to link those to a particular patient or asset, and must supply space to hold additional data provided by the RFID system. The user interfaces must provide the ability to interact with the RFID components, including input for the data that is stored in each tag and provide a means to select a particular tag (e.g., a list of asset serial numbers or patients). The need to share data among different systems introduces potential privacy concerns. However, these concerns can be addressed by ensuring that all components in the larger system comply with the facility's privacy requirements.

RFID provides a platform to expand this sharing of information and can even be used to store small subsets of a person's medical information. In one system [60] RFID tags are used to store a patient's critical care information in plain-text

to allow first-responders to quickly access this information, without having to move the patient, to provide treatment while waiting for their information to be retrieved from their insurance provider. This system is also useful in areas where there is no access to the patient’s health provider. However, this system does not address privacy issues well because the data is stored as plain-text allowing anyone to read it, but this feature is needed to provide the access for the first-responders. This example represents one key issue with privacy: how to limit access to the data from unauthorized users, while providing as many authorized users access as possible.

Confidentiality of information is another key concern for users in healthcare. Encryption is one standard means of protecting the confidentiality of information and can be used to protect the data stored in the RFID tag. While RFID tags offer limited support for high-strength cryptology, the encryption could be performed on the reader side and then the encrypted data stored on the tag. Further, higher-strength authentication and encryption methods can be employed by the backend system (e.g., database) to ensure confidentiality and availability of the information. Care must be taken to ensure that an attacker cannot infer information from the encrypted text (e.g., identify which tags have data about a cancer medication). One method of this attack is to use the Select command in the Gen-2 protocol to search user-memory on tags to identify tags with a particular bit pattern [7]. However, this attack can be prevented by using a hash function to help randomize the encrypted text that is stored in the tag for each instance.

21.5 Solutions

Tracking is a difficult problem to address in RFID because of the need to have the tag function with readers spread over a wide geographic area and controlled by a number of different institutions. However, the read ranges of most passive RFID systems are less than 5 m in free space (e.g., an open air environment) so this makes tracking someone exclusively using RFID readers difficult. Physical surveillance would most likely be a better option, because using RFID readers would require a large reader infrastructure and the attacker to have access to all of those readers (e.g., they must hack into each reader). In healthcare, access to a patient’s medical history (file) or electronic health record would provide the attacker with more information.

Confidentiality of the data stored in the RFID tag can be protected using encryption. New additions to the Gen-2 specification provide support for stronger encryption of data and for the establishment of confidential communication channels [57]. Further, ISO is developing a family of standards, ISO 29167, to standardize the deployment of security options (called “suites”) for RFID systems described in the ISO 18000 series of standards. ISO 29167-1 provides the requirements and basis for identifying the different security options; other parts in the ISO 29167 series (ISO 29167-11 defines the PRESENT-80 security suite) will define implementation of individual security options for RFID. These standards provide a unified blueprint

for developers and users to implement security options for RFID systems to meet the requirements for healthcare.

Moving the encryption and decryption operations to the backend systems, the facility's internal authentication system can be employed to prevent unauthorized access to data. The license-plate tag offers some benefits from the privacy standpoint because it contains no information other than the unique search key used during the database query process to obtain the record matching the patient (patient is linked to a unique serial number stored in the RFID tag). An attacker obtaining the tag's ID would still need to obtain access to the facility's system to access the patient's PHI which is stored in the database and not on the tag. If the attack can get into the facility's database (system) they have no need to interact with the RFID side of the system because the database contains the PHI and personal contact information for all of the facility's patients.

21.6 Conclusion

This chapter presented use cases of RFID and IoT technology in healthcare. Privacy issues associated with these technologies specific to the healthcare sector have been identified and potential solutions discussed. Tracking and confidentiality of information were the two privacy threats present in the healthcare environment. Tracking is the most difficult issue to eliminate because most mainstream commercial RFID protocols are based in part on the ability of the reader to retrieve the tag's ID number. While some methods have been proposed to address tracking RFID tags by using pseudo-IDs that change with each read (inventory) these methods require significant infrastructure and result in a closed system. In such a closed system, the RFID tag will only be able to be accessed (read) by readers within that system. This is in contrast to the IoT vision where devices, such as RFID tags, will work seamlessly with any reader to allow the RFID tag to provide maximum value or service to the user. Moving to a connected or IoT world will require systems that can support the authentication of many low-functionality devices efficiently. Physical tracking is probably a better alternative to tracking using RFID readers. Basic physical security practices, such as limited access areas (e.g., not allowing the general public into the exam areas without an escort) can prevent most tracking, either physical tracking or tracking using handheld RFID readers. Security solutions exist to address the other issues associated with RFID systems in medicine. In summary, the ability for patients to opt-out of using the RFID systems should be sufficient, as this will allow those who wise to avail themselves of the benefit to do so, it will also allow those that do not want the risk of tracking via RFID tags to accomplish that too.

The confidentiality of data stored in the RFID tag is of greater concern. The result of unauthorized access to medical data stored in the RFID tag's user-memory is a significant threat to privacy. However, this situation is easy to prevent, through the use of encryption to protect the stored data from unauthorized access. This will prevent the attacker from being able to use of interpret the data in the event that they

are able to retrieve it from the tag. Further, the new security features of the EPC Gen-2 specification and new family of ISO standards defining how to implement security and the associated security implementations (ISO 29167 family of standards) harden the RFID tags to resist unauthorized access of their user memory.

In conclusion, RFID and IoT technologies have already been used in healthcare to provide significant savings to providers, improve patient safety (e.g., ensure each patient receives the correct medications), and reduces patient waiting times have been deployed and show proven results. RFID has been used in a number of hospitals to track hand washing compliance which is one of the top ways to prevent the spread of infections. In most cases, compliance levels increased after installation of such a system. The privacy concerns identified in this chapter (tracking and confidentiality of data) can be addressed by standard solutions and procedures already employed by healthcare providers.

References

1. Department of Health, Education and Welfare. Records, Computers and the Rights of Citizens: Report of the Secretary's Advisory Committee on Automated Personal Data Systems (1973)
2. Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104–191, 110 Stat. 1936 (1996)
3. Jacques, L.B.. Electronic health records and respect for patient privacy: a prescription for compatibility. *Vanderbilt J. Entertain. Technol. Law* **13**, 441 (2010)
4. Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 (2009)
5. Federal Register. 45 CFR Parts 160 and 164 Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; Other Modifications to the HIPAA Rules; Final Rule (2013)
6. Office of the National Coordinator. Connecting Health and Care for the Nation; A Shared Nationwide Interoperability Roadmap (2014)
7. Hawrylak, P.J., Schimke, N., Hale, J., Papa, M.: Security risks associated with radio frequency identification in medical environments. *J. Med. Syst.* **36**(6), 3491–3505 (2012)
8. International Organization for Standardization: ISO/IEC DIS 18000-63 Information technology – Radio frequency identification for item management – Part 63: Parameters for air interface communications at 860 MHz to 960 MHz Type C (2013)
9. Sanders, D., Mukhi, S., Laskowski, M., Khan, M., Podaima, B.W., McLeod, R.D.: A network-enabled platform for reducing hospital emergency department waiting times using an RFID proximity location system. In: International Conference on Systems Engineering, pp. 538–543 (2008)
10. Hanada, E., Kudou, T.: Effective use of RFID in medicine. In: 2013 7th International Symposium on Medical Information and Communication Technology (ISMICT), pp. 76–80 (2013)
11. Bendavid, Y., Boeck, H., Philippe, R.: RFID-enabled traceability system for consignment and high value products: a case study in the healthcare sector. *J. Med. Syst.* **36**(6), 3473–3489 (2012)
12. Mickle, M.H., Mats, L., Hawrylak, P.J.: Physics and geometry of RFID. In: Ahson, S., Ilyas, M. (eds.) *RFID Technologies and Applications, Technology, Security, and Privacy*, pp. 3–16. CRC Press, Boca Raton (2008)

13. Hawrylak, P.J., Cain, J.T., Mickle, M.H.: RFID tags. In: Yan, L., Zhang, Y., Yang, L.T., Ning, H. (eds.) *The Internet of Things: From RFID to Pervasive Networked Systems*, pp. 1–32. Auerbach Publications, Boca Raton (2008)
14. Dehaene, W., Gielen, G., Steyaert, M., Danneels, H., Desmedt, V., De Roover, C., Li, Z., Verhelst, M., Van Helleputte, N., Radioma, S., Walravens, C., Pleysier, L.: RFID, where are they? In: Proceedings of ESSCIRC, 2009, pp. 36–43 (2009)
15. Lee, W.J., Liu, W., Chong, P.H.J., Tay, B.L.W., Leong, W.Y.: Design of applications on ultra-wideband real-time locating system. In: IEEE/ASME International Conference on Advanced Intelligent Mechatronics, 2009, pp. 1359–1364 (2009)
16. Wang, B., Toobaei, M., Danskin, R., Ngarmnil, T., Pham, L., Pham, H.: Evaluation of RFID and Wi-Fi technologies for RTLS applications in healthcare centers. In: 2013 Proceedings of PICMET '13 Technology Management in the IT-Driven Services, pp. 2690–2703 (2013)
17. Yao W., Chu, C.-H., Li, Z.: The use of RFID in healthcare: benefits and barriers. In: 2010 IEEE International Conference on RFID-Technology and Applications (RFID-TA), pp. 128–134 (2010)
18. Bhattacharya, M., Chu, C.-H., Hayya, J., Mullen, T.: An exploratory study of RFID adoption in the retail sector. *Oper. Manag. Res.* **3**(1–2), 80–89 (2010)
19. Segovis, P.: Drive savings with mobile asset management. *Health Manag. Technol.* (2012). Available: <http://www.healthmgtech.com/articles/201211/drive-savings-with-mobile-asset-management.php>
20. Kotzen, M.S.: N.J. health system saves \$1.2 million. *Health Manag. Technol.* (2013). Available: <http://www.healthmgtech.com/articles/201308/nj-health-system-saves-12-million.php>
21. Sewdberg, C.: Mexican state agency reduces donated blood wastage with RFID. *RFID J.* (2014). <http://www.rfidjournal.com/articles/view?12440>
22. Becker, E., Metsis, V., Arora, R., Vinjumur, J., Xu, Y., Makedon, F.: SmartDrawer: RFID-based smart medicine drawer for assistive environments. In: Proceedings of the 2nd International Conference on Pervasive Technologies Related To Assistive Environments (PETRA '09), pp. 1–9 (2009)
23. Vinjumur, J.K., Becker, E., Ferdous, S., Galatas, G., Makedon, F.: Web based medicine intake tracking application. In: Proceedings of the 3rd International Conference on Pervasive Technologies Related to Assistive Environments (2010)
24. Hasanuzzaman, F.M., Tian, Y.L., Liu, Q.: Identifying medicine bottles by incorporating RFID and video analysis. In: 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), pp. 528–529 (2011)
25. Rajagopalan, H., Rahmat-Samii, Y.: Ingestible RFID bio-capsule tag design for medical monitoring. In: 2010 IEEE Antennas and Propagation Society International Symposium (APSURSI), pp. 1–4 (2010)
26. Polycarpou, A.C., Dimitriou, A., Bletsas, A., Polycarpou, P.C., Papaloizou, L., Gregoriou, G., Sahalos, J.N.: On the design, installation, and evaluation of a radio-frequency identification system for healthcare applications. *IEEE Antennas Propag. Mag.* **54**(4), 255–271 (2012)
27. Castro, L., Lefebvre, E., Lefebvre, L.: Adding intelligence to mobile asset management in hospitals: the true value of RFID. *J. Med. Syst.* **37**(5), 1–17 (2013)
28. Zhao, Y., Zhou, H., Li, M.: WiTracker: an indoor positioning system based on wireless LANs. In: 2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM), pp. 1–4 (2010)
29. Au, A.W.S., Feng, C.; Valaee, S., Reyes, S., Sorour, S., Markowitz, S.N., Gold, D., Gordon, K., Eizenman, M.: Indoor tracking and navigation using received signal strength and compressive sensing on a mobile device. *IEEE Trans. Mob. Comput.* **12**(10), 2050–2062 (2013)
30. Zhang, D., Zhou, J., Guo, M., Cao, J., Li, T.: TASA: tag-free activity sensing using RFID tag arrays. *IEEE Trans. Parallel Distrib. Syst.* **22**(4), 558–570 (2011)
31. Chen, R.-C., Lin, Y.-H.: Apply Kalman filter to RFID Received Signal Strength processing for indoor location. In: 2012 4th International Conference on Awareness Science and Technology (iCAST), pp. 73–77, 21–24 (2012)

32. Zhao, J., Zhang, Y., Ye, M.: Research on the received signal strength indication location algorithm for RFID system. In: International Symposium on Communications and Information Technologies, 2006. ISCIT '06, pp. 881–885 (2006)
33. Huang, Y., Brennan, P.V., Seeds, A.: Active RFID location system based on time-difference measurement using a linear FM chirp tag signal. In: IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008, pp. 1–5 (2008)
34. Zou, Z., Deng, T., Zou, Q., Sarmiento, M.D., Jonsson, F., Zheng, L.-R.: Energy detection receiver with TOA estimation enabling positioning in passive UWB-RFID system. In: 2010 IEEE International Conference on Ultra-Wideband (ICUWB), vol. 2, pp. 1–4 (2010)
35. Zhai, C., Zou, Z., Zhou, Q., Zheng, L.: A software defined radio platform for passive UWB-RFID localization. In: 2012 IEEE International Conference on Wireless Information Technology and Systems (ICWITS), pp. 1–4 (2012)
36. Ai, Z., Liu, Y.: Research on the TDOA measurement of active RFID real time location system. In: 2010 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), vol. 2, pp. 410–412 (2010)
37. Azzouzi, S., Cremer, M., Dettmar, U., Kronberger, R., Knie, T.: New measurement results for the localization of UHF RFID transponders using an Angle of Arrival (AoA) approach. In: 2011 IEEE International Conference on RFID, pp. 91–97 (2011)
38. Hua, M.-C., Peng, G.-C., Lai, Y.J., Liu, H.-C.: Angle of arrival estimation for passive UHF RFID tag backscatter signal. In: IEEE International Conference on and IEEE Cyber, Physical and Social Computing Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), pp. 1865–1869 (2013)
39. Toplan, E., Ersoy, C.: RFID based indoor location determination for elderly tracking. In: 20th Signal Processing and Communications Applications Conference (SIU), pp. 1–4 (2012)
40. Sutherland, J., van den Heuvel, W.-J.: Towards an intelligent hospital environment: adaptive workflow in the OR of the future. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences, 2006, HICSS '06, vol. 5, pp. 100b (2006). doi:10.1109/HICSS.2006.494
41. Okoniewska, B., Graham, A., Gavrilova, M., Wah, D., Gilgen, J., Coke, J., Burden, J., Nayyar, S., Kaunda, J., Yergens, D., Baylis, B., Ghali, W.A.: Multidimensional evaluation of a radio frequency identification Wi-Fi location tracking system in an acute-care hospital setting. *J. Am. Med. Inform. Assoc.* **19**(4), 674–679 (2012)
42. Arcega, L., Font, J., Cetina, C.: Towards memory-aware services and browsing through lifelogging sensing. *Sensors* **13**(11), 15113–15137 (2013)
43. Blasco, R., Marco, Á., Casas, R., Cirujano, D., Picking, R.: A smart kitchen for ambient assisted living. *Sensors* **14**(1), 1629–1653 (2014)
44. Wang, S.-W., Chen, W.-H., Ong, C.-S., Liu, L., Chuang, Y.-W.: RFID application in hospitals: a case study on a demonstration RFID project in a Taiwan hospital. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences, vol. 8, pp. 184a (2006)
45. Hanser, F., Gruenerbl, A., Rodegast, C., Lukowicz, P.: Design and real life deployment of a pervasive monitoring system for dementia patients. In: Second International Conference on Pervasive Computing Technologies for Healthcare, pp. 279–280 (2008)
46. Lee, S.-Y., Cho, G.-S.: A simulation study for the operations analysis of dynamic planning in container terminals considering RTLS. In: Second International Conference on Innovative Computing, Information and Control (ICICIC '07), pp. 116 (2007)
47. Cangialosi, A., Monaly, J.E., Yang, S.C.: Leveraging RFID in hospitals: patient life cycle and mobility perspectives. *IEEE Commun. Mag.* **45**(9), 18–23 (2007)
48. Xiong, J., Seet, B.-C., Symonds, J.: Human activity inference for ubiquitous RFID-based applications. In: 2009 Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, pp. 304–309 (2009)
49. Saygin, C.: Adaptive inventory management using RFID data. *Int. J. Adv. Manuf. Technol.* **32**(9–10), 1045–1051 (2007)
50. Goebel, C., Günther, O.: Benchmarking RFID profitability in complex retail distribution systems. *Electron. Mark.* **19**(2–3), 103–114 (2009)

51. Bustillo, M.: Wal-Mart radio tags to track clothing. *Wall Street J.* July 23, 2010. <http://www.wsj.com/articles/SB10001424052748704421304575383213061198090>
52. Shieh, H.-L., Lin, S.-F., Chang, W.-S.: RFID medicine management system. In: 2012 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 5, pp. 1890–1894 (2012)
53. Juels, A.: 2006. RFID security and privacy: a research survey. *IEEE J. Sel. Areas Commun.* **24**(2), 381–394 (2006)
54. Garfinkel, S.L., Juels, A., Pappu, R.: RFID privacy: an overview of problems and proposed solutions. *IEEE Secur. Priv.* **3**(3), 34–43 (2005)
55. Lee, Y.K., Batina, L., Singelée, D., Verbauwhede, I.: Low-cost untraceable authentication protocols for RFID. In: Proceedings of the Third ACM Conference on Wireless Network Security (WiSec ‘10), pp. 55–64 (2010)
56. Li, Y., Teraoka, F.: Privacy protection for low-cost RFID tags in IoT systems. In: Proceedings of the 7th International Conference on Future Internet Technologies (CFI ‘12), pp. 60–65 (2012)
57. Engels, D.W., Kang, Y.S., Wang, J.: On security with the new Gen2 RFID security framework. In: 2013 IEEE International Conference on RFID, pp. 144–151 (2013)
58. Karjoth, G., Moskowitz, P.A.: Disabling RFID tags with visible confirmation: clipped tags are silenced. In: Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society (WPES ‘05), pp. 27–30 (2005)
59. Barlow, R.: Next-generation tracking: go beyond tracking people, products and equipment. *Health Manag. Technol.* **35**(10), 6–11 (2014)
60. Hart, C., Hawrylak, P.J.: Using radio frequency identification (RFID) tags to store medical information needed by first responders: data format, privacy, and security. *Int. J. Comput. Methods Algorithms Med.* **3**(3), 10–26 (2012)

Chapter 22

Privacy Preserving Classification of ECG Signals in Mobile e-Health Applications

Riccardo Lazzeretti and Mauro Barni

Abstract Privacy protection is an emerging problem in mobile Health applications. On one hand, cloud services enable to store personal medical data, making them always available, and providing preliminary analysis on them, on the other hand, storing personal health data entails serious threats to users privacy. Privacy preserving solutions, such as Secure Multi-Party Computation techniques, give to non-trusted parties the opportunity of processing biomedical signals while encrypted. This chapter focuses on the development of a privacy preserving automatic diagnosis system whereby a remote server classifies an ElectroCardioGram (ECG) signal provided by the client without obtaining neither any information about the signal itself, nor the final result of the classification. Specifically, we present and compare three secure implementations of ECG classifiers: Linear Branching Programs (a particular kind of decision tree) with Quadratic Discriminant Functions, Linear Branching Programs with Linear Discriminant Functions and Neural Networks. Moreover we describe a protocol that permits to evaluate the quality of an encrypted ECG. The chapter provides a signal processing analysis aiming at satisfying both accuracy and complexity requirements. The described systems prove that carrying out complex tasks like ECG classification in the encrypted domain is indeed possible in the semi-honest model, paving the way to interesting future applications wherein privacy of signal owners is protected by applying high security standards.

22.1 Introduction

The healthcare industry is on the edge of an unprecedented revolution made possible by the widespread diffusion of Information and Communication Technology (ICT). ICT, in fact, has a lot to offer in terms of personalized online self-service, medical error reduction, customer data collection and more. Mobile healthcare applications based on e-cloud services and remote data storage systems, add a further very promising perspective to this on-going revolution. Yet, concerns regarding the

R. Lazzeretti (✉) • M. Barni

Department of Information Engineering & Mathematics, University of Siena, Siena, Italy
e-mail: lazzeretti@diism.unisi.it; barni@di.unisi.it

privacy of the users of e-health services risk to slow down or even inhibit the diffusion of these new services, causing a huge societal and economical damage.

Storing and distributing the biomedical signals and other privacy-sensitive data in encrypted form is an obvious solution to the above concerns, which, however, comes at the price of reduced flexibility and efficiency of e-health services. If biomedical signals and data were available only in encrypted form, functionalities like content-based queries or signal analysis for remote diagnosis could not be easily implemented thus depriving e-health systems of some of their most appealing features.

The possibility of processing biomedical signals while they are encrypted would clearly provide a secure and elegant way to overcome the above problems. Such an apparently unfeasible task is indeed possible thanks to the possibilities offered by s.p.e.d. (signal processing in the encrypted domain) technology [18, 33], whereby it is possible to process biomedical and other data without revealing the content of the data, and even the processing results, to non-intended users. A simple, yet rather general, scheme for s.p.e.d. applications, encompasses the presence of two players: one party, hereafter referred to as the Client (\mathcal{C}) owns a signal that has to be processed in some way by the other party, hereafter referred to as the Server (\mathcal{S}). Since \mathcal{C} and \mathcal{S} do not trust each other, \mathcal{S} must process the signal owned by \mathcal{C} without getting any information about it, not even the result of the processing. At the same time, \mathcal{S} wants to protect the information it uses to process the signal, since this represents the basis of the service it offers. More specifically, given a general functionality $f()$, whose specific instantiation is defined by a set of parameters p and an input signal x , the computation of $f(x; p)$ in a s.p.e.d. framework requires that \mathcal{S} , who knows p , is able to compute the result of $f(x; p)$, without getting any information about x and the result of the computation. At the same time, \mathcal{C} should get no information regarding p .

From a technical point of view, s.p.e.d. protocols may adopt two different approaches based, respectively, on Homomorphic Encryption (HE) [20, 45] and Secure Multi-Party Computation (SMPc) [25, 43].¹ The former approach relies on the particular homomorphic properties of some cryptosystems, for which one operation on a pair of non-encrypted values (e.g. an addition) corresponds to a simple operation (e.g. a multiplication) on the encrypted messages. For example, by using Paillier cryptosystem [39], it is possible to evaluate the sum of encrypted values without decrypting them. This allows to evaluate any linear operation on the server's side, while other operations can be implemented by interacting with the key owner. The STPC approach to s.p.e.d. relies on the seminal works by Goldreich [19] and Yao [49, 50], in which the possibility of computing any functionality described by an acyclic Boolean circuit on private data was demonstrated. By using Yao's Garbled Circuits, two parties can jointly evaluate a boolean circuit. The first party “encrypts” the circuit and sends his input bits in encrypted form to the other party, which in turn “decrypts” the circuit gate by gate by using his encrypted inputs

¹When only two players are involved, SMPc reduces to Secure Two Party Computation (STPC).

together with the received ones as decryption keys. The complexity of the protocols depends on the functionality to be evaluated, sometimes making the Homomorphic Encryption preferable to Garbled Circuit, or vice versa. Hybrid solutions, exploiting the best characteristics of the two approaches, have been proposed as well [30, 32]. The idea is to split the protocol in subprotocols and evaluate each of them with the technique having the lowest complexity. Then the blocks are connected by using interfaces based on additive blinding.

In this chapter we describe a specific application in which s.p.e.d. technology is used for the development of a privacy preserving automatic diagnosis system whereby a remote server classifies an ElectroCardioGram (ECG) signal provided by the client without obtaining any information about the signal itself and the final result of the classification. Specifically, we present and compare the secure implementation of three different ECG classifiers, based respectively on Linear Branching Programs (a particular kind of decision tree) with Quadratic Discriminant Functions [4], Neural Networks [7] and Linear Branching Programs with Linear Discriminant Functions, here addressed for the first time. In order to allow a fair comparison among the different implementations, all of them rely on the same features. In addition to the design of an accurate and efficient ECG classification algorithm, the deployment of reliable remote healthcare solutions requires that possible errors made by remote patients, who can not rely on the assistance of expert medical personnel, are corrected or at least detected. In the case of ECG classification, for instance, it is possible that the user connects the electrodes in a wrong way, thus sending to the server a poor quality signal. This, in turn, increases the probability of a classification error. For this reason, we end this chapter by presenting a protocol that allows the evaluation of the quality of ECG signals by working on encrypted data.

Rather than providing full-fledged solutions, the goal of this chapter is to show that reliable ECG classification in the encrypted domain is indeed possible. Particular attention is given to describe the trade-off between classification accuracy and efficiency. As a matter of fact, the central question in s.p.e.d. technology is not *whether* a certain functionality can be computed, but *how efficiently* this can be done. As it will be clear from the subsequent sections, designing an efficient s.p.e.d. protocol requires that the representation of the input and intermediate values of the computation is carefully studied to simplify the protocol without sacrificing the accuracy of the computation. In a similar way, the classification algorithm must be simplified as much as possible, by reducing the number of operations whose implementation in s.p.e.d. setting is most problematic.

The chapter is organized as follows. In Sect. 22.2, the exact ECG classification problem we face with is defined and the plain version of the classification algorithm the s.p.e.d. protocols rely on is described. Section 22.3, introduces the basic cryptographic primitives at the basis of any s.p.e.d. protocol. The description focuses more on the usage and functional properties of the primitives rather than on their mathematical definition. The security model used throughout the chapter is presented as well. Section 22.4 describes a s.p.e.d. implementation of Linear Branching Programs (LBP) which will be the basis for the first two

ECG classifiers presented in the chapter, namely a LBP classifier based on a quadratic discriminant function (Sect. 22.4.2) and a LBP classifier based on a linear discriminant function (Sect. 22.4.3). Section 22.5 presents the third classifier, based on a s.p.e.d. implementation of Neural Networks (NN). Section 22.6 describes a protocol for the evaluation of the quality of encrypted ECG signals. The chapter ends with some conclusions in Sect. 22.7.

22.2 Plain Protocol

Before describing the s.p.e.d. implementation of three different ECG classifiers, we present the plain domain algorithm for ECG classification at the basis of the s.p.e.d. protocols.

The goal is to classify single heart beats according to six classes: Normal Sinus rhythm (NSR), Atrial Premature Contractions (APC), Premature Ventricular Contractions (PVC), Ventricular Fibrillation (VF), Ventricular Tachycardia (VT) and SupraVentricular Tachycardia (SVT). The classification algorithm that inspired the protocols described here derives from the system introduced by Ge et al. ([21] and [1, Chap. 8]). Specifically, it relies on a rather general technique based on Autoregressive (AR) models for ECG description and a subsequent Quadratic Discriminant Function (QDF) classifier where each ECG interval corresponding to one heart beat is modeled by means of a fourth order AR model.

The choice of the algorithm is justified first of all by the good classification accuracy it ensures, secondly because it can be easily implemented in a s.p.e.d. framework. The overall architecture of the classifier is summarized by the block diagram in Fig. 22.1. The input of the system is a filtered ECG chunk composed by 300 samples surrounding the R peak (the highest peaks of a heart beat signal). The samples are modeled by a fourth order AR model whose parameters, together with two additional parameters accounting for the modeling error, are used by the QDF block to produce the feature vector that will be used in the subsequent classification. Each block of Fig. 22.1 is described in the following.

Autoregressive Modeling: An AR model [12] is a linear predictor that attempts to estimate the value of a sample s_n from the previous p samples ($s_{n-1}, s_{n-2}, \dots, s_{n-p}$), through an equation having the following form:

$$s_n = \sum_{i=1}^p a_i s_{n-i} + \epsilon_n \quad (22.1)$$

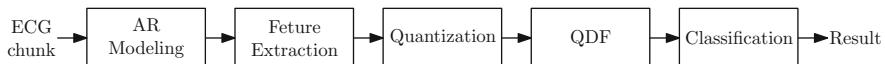


Fig. 22.1 Block diagram

where $\{a_i\}_{i=1..p}$ are the AR coefficients, $\{\epsilon_i\}_{i=1..N}$ is the prediction error sequence and p is the model degree. A good AR model chooses the AR coefficients which minimize the mean square error between the original sequence and the one predicted by the model. As shown in [21], for an accurate description of ECG signals an AR model of order $p = 4$ is sufficient.

Feature Extraction: Six features are extracted from the AR model. The feature vector is the following:

$$\mathbf{f} = (f_1, f_2, f_3, f_4, f_5, f_6)^T = (a_1, a_2, a_3, a_4, n_1, n_2)^T \quad (22.2)$$

The first four features are the coefficients of the AR model; the last two features measure the quality of the predictor by counting the number of samples wherein the amplitude of the AR model respectively exceeds or not exceed an empiric threshold set to $th = 0.25 \max_n (|\epsilon_n|)$, hence n_1 is the number of times that $|\epsilon_n| < th$ and n_2 is the number of times that $|\epsilon_n| > th$.

Quadratic Discriminant Function: With QDF classifications, the classifier operates on a composite feature vector \mathbf{x} instead than the feature vector \mathbf{f} . The composite vector contains the features in \mathbf{f} , their square values and their cross products, namely:

$$\begin{aligned} \mathbf{x} &= (1, f_1, \dots, f_6, f_1^2, \dots, f_6^2, f_1 f_2, \dots, f_1 f_6, f_2 f_3, \dots, f_2 f_6, \dots, f_5 f_6)^T \\ &= (x_1, x_2, \dots, x_{28})^T \end{aligned} \quad (22.3)$$

The goal of this step is to project the vector \mathbf{x} onto six directions \mathbf{w}_i , obtaining a 6-long vector $\mathbf{y} = \mathbf{W}\mathbf{x}$, whose elements are used in the classification tree described in the next step. The sign of each y_i value identifies the path that has to be chosen in the relative node. The rows of matrix \mathbf{W} are the vectors \mathbf{w}_i . The matrix \mathbf{W} contains part of the knowledge embedded within the classification system, and is computed by relying on a set of training ECG's. In particular, the matrix \mathbf{W} is computed by minimizing the mean square error between the actual vector \mathbf{y} and the target values of \mathbf{y} . A vector used for the training set is said \mathbf{y}_t . The desired output for each of the diseases is given in Table 22.1.

Table 22.1 Classification pattern (“*” means that the value of the variable does not influence the classification)

y_1	y_2	y_3	y_4	y_5	y_6	Disease
1	1	*	1	1	*	NSR
1	1	*	1	*	1	NSR
1	1	*	1	-1	*	APC
1	1	*	1	*	-1	PVC
-1	*	-1	*	*	*	VF
-1	*	1	*	*	*	VT
1	-1	*	*	*	*	SVT

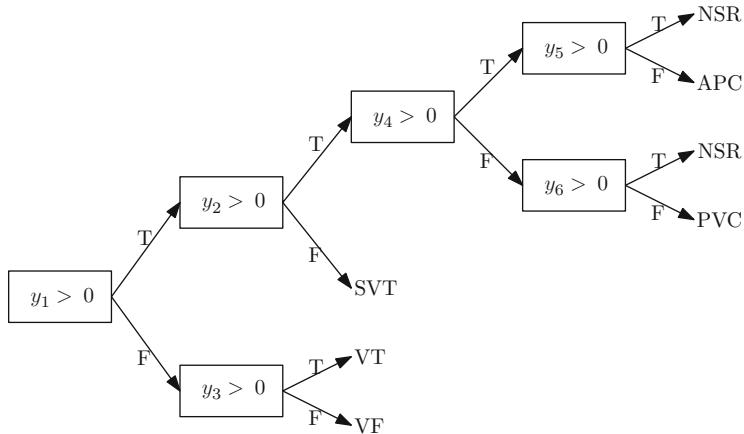


Fig. 22.2 The decision graph leading to ECG segment classification. Given the array y_1, \dots, y_6 , the tree is traversed according to the result of the comparison of the values with 0 in each node, following the true (T) or false (F) edges

Let us now consider the estimation of a single row of \mathbf{W} , that is \mathbf{w}_i . During the training, for each ECG segment j , the composite feature vector \mathbf{f}_j^c is computed. Assuming that D is the number of ECG segments used for the training of a particular \mathbf{w}_i , the vector that minimizes the error between the target output and the actual one is given by

$$\mathbf{w}_i = (\mathbf{X}_i^T \mathbf{X}_i)^{-1} \mathbf{X}_i^T \mathbf{y}_{ti}. \quad (22.4)$$

where $\mathbf{X}_i = (\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_D)$ is a $D \times 28$ matrix containing the composite feature vectors of the training set, and where $\mathbf{y}_{ti} = (y_{i1}, \dots, y_{iD})^T$ is the column vector of the i th component of the target vector responses ($y_{ij} \in \{-1, +1\}$).

Classification: To classify an ECG segment, the output \mathbf{y} of the QDF is used in the nodes of the classification tree given in Fig. 22.2, whose structure, presented in [1, Chap. 8], is designed to better separate the classes. The structure of the tree depends on the fact that there is a multiple dichotomy of six classes of samples. The Euclidean center distance between these classes were computed for determining the groupings of classes at each stage, obtaining that APC/NSR/PVC, VT/VF and SVT form one group respectively due to small values of the Euclidean center distance within the same group and large values between different groups.

The path along the tree is chosen according the signs of the y_i values. The first node separates VT/VF ($y_1 < 0$) from APC/NSR/PVC and SVT ($y_1 > 0$). Similarly, in the second node (based on y_2), VT and VF were differentiated. In the third stage (y_3), SVT was distinguished from NSR, APC and PVC. In the later stages ($y_4; y_5; y_6$), NSR, APV and PVC were distinguished from each other and classified.

Table 22.2 QDF-based classification results (results from [1, Chap. 8])

Testing data set		NSR	APC	PVC	SVT	VT	VF	Accuracy result (%)
140	NSR	135	3	1	0	1	0	96.4
140	APC	8	131	0	1	0	0	93.5
140	PVC	4	0	133	2	1	0	95.0
140	SVT	0	1	2	137	0	0	97.9
140	VT	0	0	0	0	136	4	97.1
140	VF	0	0	1	0	4	135	96.4

Table 22.3 QDF-based classification results obtained in the tests

Testing data set		NSR	APC	PVC	SVT	VT	VF	Accuracy result (%)
140	NSR	132	8	0	0	0	0	94.3
140	APC	22	118	0	0	0	0	84.3
140	PVC	0	2	138	0	0	0	98.6
140	SVT	0	0	0	140	0	0	100.0
140	VT	0	0	0	0	100	40	71.4
140	VF	0	0	0	0	26	114	81.4

22.2.1 Classification Results

The classification results on test data provided by Ge et al. in [1, Chap. 8] are given in Table 22.2. The authors selected 200 heart beats for each class and used 60 of them for training, the others for testing. The mean classification accuracy is 96.1 %.

We re-implemented the protocol proposed by Ge et al. and, since the dataset used in the original paper was not available, a different dataset has been used in the experiments. The new dataset was still obtained from PhysioBank [24]. The dataset is built by selecting 200 samples of each class from 1 or at most 2 patterns. From these samples, 60 samples are selected at random for each class and used to build the training set. The others have been used for the test set. The performance we obtained on the new dataset are lower than those given in [21], the goal here, however, is only to demonstrate that a s.p.e.d. implementation of the protocol can provide results comparable to a plain implementation with a reasonable complexity. The algorithm developed was able to classify correctly 88.3 % of the test set. Table 22.3 shows the corresponding confusion matrix. In the rest of the chapter we will always refer to this table.

22.3 Cryptographic Primitives

In this section, we present the cryptographic primitives behind the proposed protocols, namely Homomorphic Encryption (HE) [45] and Garbled Circuits (GC)

[49]. Also Oblivious Transfer (OT) [19] is presented, being an important component of the GC tool. We conclude the section by presenting how HE and GC can be connected to develop more efficient protocols.

Before describing the tools, we discuss the security model adopted throughout the chapter. The *semi-honest* security model [25, Chap. 3.1] is considered here, according to which the involved parties are assumed to follow the protocol but try to learn as much as possible from the exchanged messages and their private inputs. Despite its simplicity, designing and evaluating the performance of protocols in the semi-honest model is a first stepping stone towards the development of protocols with stronger security guarantees. The security of the Paillier HE, here considered, and GC in the semi-honest setting are demonstrated in [39] and [38] respectively, while the security of their composition in hybrid protocols is proven in [30].

The complexity of privacy preserving protocols depend on many factors. Computational complexity is of course the most interesting measure, however it depends on the architecture of the devices used for the computation and is usually measured as the CPU runtime. A more abstract measure can be provided by counting the number of the most expensive operations. On the other hand, the communication complexity of different solutions can be easily compared by measuring (or estimating) the bandwidth required by each protocol. The number of communication rounds required by the protocols is also an important quantity, since each transmission round introduces a delay due to network latency.

22.3.1 Homomorphic Encryption

With an additively homomorphic, asymmetric, encryption scheme, it is possible to obtain the encryption of the sum of two values a and b available in encrypted form through another operation \boxplus evaluated on the corresponding ciphertexts. In other words, by denoting with $\llbracket \cdot \rrbracket$ the encryption operator, we have $\llbracket a + b \rrbracket = \llbracket a \rrbracket \boxplus \llbracket b \rrbracket$, where equality indicates that the decryption of $\llbracket a + b \rrbracket$ and $\llbracket a \rrbracket \boxplus \llbracket b \rrbracket$ results in the same plain value $a + b$. Considering that the product between two values, one of them available in non-encrypted form, can be implemented by repeated sums, it is generally possible to implement the product between an encrypted value and a known integer quantity by means of an operation \boxtimes , i.e. $\llbracket ab \rrbracket = \llbracket a \rrbracket \boxtimes b$.

The most widely used additively homomorphic cryptosystem is Paillier cryptosystem [39] with plaintext space \mathbb{Z}_N and ciphertext space $\mathbb{Z}_{N^2}^*$, where N is a T -bit RSA modulus and a ciphertext is represented with $2T$ bits. For short term security $T = 1248$ [23]. By using Paillier cryptosystem, addition and product are mapped respectively into product and exponentiation, i.e. $\llbracket a + b \rrbracket = \llbracket a \rrbracket \cdot \llbracket b \rrbracket \bmod N^2$ and $\llbracket ab \rrbracket = \llbracket a \rrbracket^b \bmod N^2$. More complex functionalities, such as bit decomposition [47] and comparison [15], can be evaluated by interacting with the owner of the decryption key. The communication complexity of HE-based protocols is mainly related to the number of ciphertexts to be transmitted and the number of rounds

necessary for the evaluation of the protocol. The computational complexity is usually measured in terms of number of modular exponentiations, where encryption and decryption complexities are assumed to be similar to exponentiations.

Multiplicative homomorphic cryptosystems exist as well [17, 44], allowing the evaluation of products between encrypted values, but they have a lower practical utility with respect to additive HE.

Fully Homomorphic Encryption schemes allow both the evaluation of additions and products in the encrypted domain. C. Gentry [22] introduced the first secure *Somewhat Homomorphic Encryption* (SHE) and *Fully Homomorphic Encryption* (FHE) schemes, working on binary data. SHE allows the evaluation of a limited number of additions and multiplications, while FHE extends SHE to bypass such a restriction by performing several re-encryptions of the encrypted values during the computation at the price of a huge increment of memory and computational complexity, thus making all FHE schemes proposed so far highly impractical. Actually, as shown in [14], the key generation with small security needs 36 s and the public key size is 9.6 MB, while with large security key generation it needs 43 min and the public key is 802 MB long. Re-encryption needs 10.5 s with short security and more than 14 min with long security. Although key generation and transmission are really expensive, they are executed only once and their complexity can be acceptable, on the other side re-encryption is frequently performed, making secure protocols slow. By using the original Gentry's SHE scheme and subsequent improvements, it is possible to evaluate binary circuits composed by up to a maximum number of XOR and AND gates directly on \mathcal{S} 's side without interacting with \mathcal{C} , thus making protocols based on SHE very appealing for clients equipped with low power devices. Efficient SHE solutions can be designed to evaluate circuits having a given (small) number of AND gates and then transformed into more expensive FHE solutions, if necessary. Luckily in most s.p.e.d. protocols, the number of required operations is known in advance, thus making the use of protocols based on SHE possible.

A further simplification has been introduced in [42], where a SHE scheme operating on integer values is introduced, thus allowing to encrypt each input directly, instead of decomposing it into bits and then use bitwise encryption. On the negative side, SHE (or FHE) schemes working on integers permit only the evaluation of polynomial functions (up to a certain degree for SHE).

22.3.2 Oblivious Transfer

An Oblivious Transfer (OT) protocol [19] allows one party, the chooser, to select one out of two (or more) inputs provided by another party, the sender, in a way that protects both parties: the sender is assured that the chooser does not receive more information than it is entitled, while the chooser is assured that the sender does not come to know which input he received.

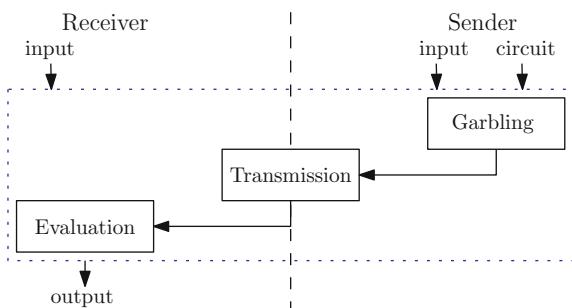
OT protocols consist of two phases: an offline and an online phase. It is customary to move the set up operations and a great part of the most computationally expensive operations to the offline phase, which is performed during inactivity periods, and during which the chooser and the sender run several OT protocols computed on random values. In such a phase, a great number of OTs can be evaluated in parallel by using the OT extension presented in [3], which permits to evaluate up to 700,000 OTs per second over a common wi-fi link. During the online phase, the results of the precomputed OTs are used to update the OTs to the actual values [8], through the transmission of only $2m$ bits in 2 rounds for each OT [27], where m is the message bitlength.

We indicate with OT_m^k the parallel execution of k OT protocols operating on m -bits long messages.

22.3.3 Garbled Circuits

Any boolean circuit containing no cycle can be privately evaluated on secret inputs by using Yao's Garbled Circuits (GC) [49, 50]. Despite several optimizations proposed later, the overall protocol for GC evaluation is still similar to the first one proposed by Yao. As shown in Fig. 22.3, a GC is evaluated in three steps. During the garbling step, one party, the sender, associates a couple of secrets (one for each logical value) to each wire of the circuit and garbles each gate by encrypting, for each row of the corresponding truth table, the secret associated to the output by using the two secrets associated to the inputs. In the transmission phase, the sender transmits the garbled tables to the other party, the receiver. Moreover the sender transmits the secrets associated to its input wires, while the receiver obtains the secrets of his input wires by means of an OT protocol. Finally, during the evaluation phase, the receiver decrypts the secrets, gate by gate, starting from the gates connected to the inputs (for a detailed description of each of the above steps we refer to [36]).

Fig. 22.3 Garbled circuit scheme



Thanks to a number of recently proposed optimizations, the complexity associated to the garbling, transmission and evaluation of XOR gates can be neglected [31], while for each non-XOR binary gate, thanks to the Garbled Row Reduction scheme [41], binary gate garbling requires the computation of 3 Hash functions and the transmission of $3t$ bits, where t is a security parameter ($t = 80$ for short term security). In addition, gate evaluation requires the computation of a Hash function with probability $3/4$.² With regard to the transmission of input secrets, for each input bit of the sender, a secret of t bits is transmitted, while for each input bit of the receiver an OT is evaluated ($2t$ bits transmitted online). By considering that the values associated with the wires are correlated [31], additional bandwidth in the OT extension protocol can be saved as shown in [3]. Other optimizations can be obtained by observing that the gates of the circuit can be topologically sorted to pipeline the circuit generation and transmission [26], and by using fixed-key AES in gate garbling [9].

We underline that if the sender and the receiver know in advance the to-be-evaluated function, garbling and circuit transmission can be performed offline, when the inputs are not yet available.

22.3.4 Hybrid Protocols

The use of hybrid protocols permits to efficiently evaluate functionalities for which full-HE or full-GC solutions would not be efficient (or even impossible). Given that GC and HE rely on different ways of representing data, conversion from homomorphic ciphertexts to garbled secrets (or vice versa) must be performed by resorting to interfacing protocols based on additive blinding. In particular, by referring to the protocols described in [32], it is easy to derive that the conversion of an ℓ -bit value from HE to GC requires the on-line transmission of additional $2T + 7\ell t$ bits, while the conversion from GC to HE requires an overhead of $2T + (\ell + \tau)5t$ bits, where τ is an obfuscation security parameter (usually $\tau = 80$). Hybrid protocols have been used to develop efficient solutions for biometric identification [11, 13, 46], biomedical analysis [4, 6, 7, 37], secure approximated general function evaluation [40], etc. have been proposed. Other cryptographic primitives can also be used in hybrid protocols. For example, in [16], the authors present a framework, called ABY, that efficiently combines secure computation schemes based on Arithmetic sharing, Boolean sharing, and Yao's garbled circuits.

²The garbling of a d -input gate requires the computation of $2^d - 1$ Hash functions and the transmission of $(2^d - 1)t$ bits, while gate evaluation requires the computation of a Hash function with probability $1 - 1/2^d$.

22.4 Privacy Preserving Linear Branching Program

In this section, we formally introduce Linear Branching Programs (LBP) [7], and show how they can be implemented in a s.p.e.d. framework [4, 5, 7]. We then introduced two different ECG classifiers based on LBP.

22.4.1 Linear Branching Programs (LBP)

A Branching Program (BP) is a classification tool based on decision graph that can be easily implemented by relying on STPC tools [29]. In BP, starting from the root, in each node the left or right sub-tree is chosen according to the value of a discriminant feature, until a classification leaf is reached. A Linear Branching Program (LBP), such as the one shown in Fig. 22.2, is an evolution of standard BP, where the decision is not performed according to a single feature, but according to the result of a scalar product (a linear operation) between an array of features and an array of discriminant features. Such solution is useful to separate elements of different classes in more steps, where the first steps distinguish among groups of classes and then the choice of the classification class is performed in the leaves. A formal definition of LBP is given below.

Definition 22.1 (Linear Branching Program). Let $\mathbf{x}^{(\ell)} = [x_1^{(\ell)}, \dots, x_n^{(\ell)}]^T$ be the attribute vector of signed ℓ -bit integer values. A binary **Linear Branching Program (LBP)** \mathcal{L} is a triple $\langle \{P_1, \dots, P_z\}, Left, Right \rangle$. The first element is a set of z nodes consisting of d “decision nodes” P_1, \dots, P_d followed by $z - d$ “classification nodes” P_{d+1}, \dots, P_z .

Decision nodes P_i , $1 \leq i \leq d$ are the internal nodes of the LBP. Each $P_i := \langle \mathbf{w}_i^{(\ell)}, t_i^{(\ell')} \rangle$ is a pair, where $\mathbf{w}_i^{(\ell)} = [w_{i,1}^{(\ell)}, \dots, w_{i,n}^{(\ell)}]$ is the linear combination vector consisting of n signed ℓ -bit integer values (“weights”) and $t_i^{(\ell')}$ is the signed ℓ' -bit integer value (“threshold”) with which $\mathbf{w}_i^{(\ell)} \mathbf{x}^{(\ell)} = \sum_{j=1}^n w_{i,j}^{(\ell)} x_j^{(\ell)}$ is compared in the node. $Left(i)$ is the index of the next node if $\mathbf{w}_i^{(\ell)} \mathbf{x}^{(\ell)} \leq t_i^{(\ell')}$; $Right(i)$ is the index of the next node if $\mathbf{w}_i^{(\ell)} \mathbf{x}^{(\ell)} > t_i^{(\ell')}$. Functions $Left()$ and $Right()$ are such that the resulting directed graph is acyclic.

Classification nodes $P_j := \langle c_j \rangle$, $j \in \{d + 1, \dots, z\}$ are the leaf nodes of the LBP consisting of a single classification label c_j each.

Given a LBP \mathcal{L} and an input vector $\mathbf{x}^{(\ell)}$, the evaluation starts from the root node P_1 . If $\mathbf{w}_1^{(\ell)} \mathbf{x}^{(\ell)} \leq t_1^{(\ell')}$, we move to decision node indicated by $Left(1)$, else to $Right(1)$. Repeat this process recursively (with corresponding $\mathbf{w}_i^{(\ell)}$ and $(t_i^{(\ell')})$), one of the classification nodes is reached, obtaining the classification $c = \mathcal{L}(\mathbf{x}^{(\ell)})$.

In the general case of LBPs, the bit-length ℓ' of the threshold values $t_i^{(\ell')}$ has to be chosen according a worst case analysis aiming to determinate the maximum value of the linear combinations [4], obtaining $\ell' = 2\ell + \lceil \log_2 n \rceil - 1$.

In the following, we describe two possible s.p.e.d. implementations of a generic LBP: a full-GC protocol and a hybrid protocol.

22.4.1.1 Full-GC Implementation

We here describe a boolean circuit implementing LBP that can be evaluated by using GC. First, \mathcal{S} garbles a boolean circuit C implementing the LBP. Such circuit has ℓ -bit inputs $x_1^{(\ell)}, \dots, x_n^{(\ell)}$, provided by \mathcal{C} and ℓ -bit input $w_{i,j}^{(\ell)}$ and ℓ' -bit \mathcal{S} 's input $t_i^{(\ell)}$, with $i \in \{1, \dots, d\}$, $j \in \{1, \dots, n\}$, provided by \mathcal{S} . The output of the circuit is composed by the bits $\sigma_1, \dots, \sigma_d$ that obliviously computes the intended functionality as described below.

The circuit is composed by d sub-circuits, one for each node and a 6-input gate returning the final classification. For each node i , with $1 \leq i \leq d$, the circuit shown in Fig. 22.4 implements $\sigma_i = (\mathbf{w}_i^{(\ell)} \mathbf{x}^{(\ell)} > t_i^{(\ell)}) = (\sum_{j=1}^n w_{i,j}^{(\ell)} \cdot x_j^{(\ell)} > t_i^{(\ell)})$. First of all the magnitudes of the feature elements $x_j^{(\ell)}$ are multiplied by the magnitude

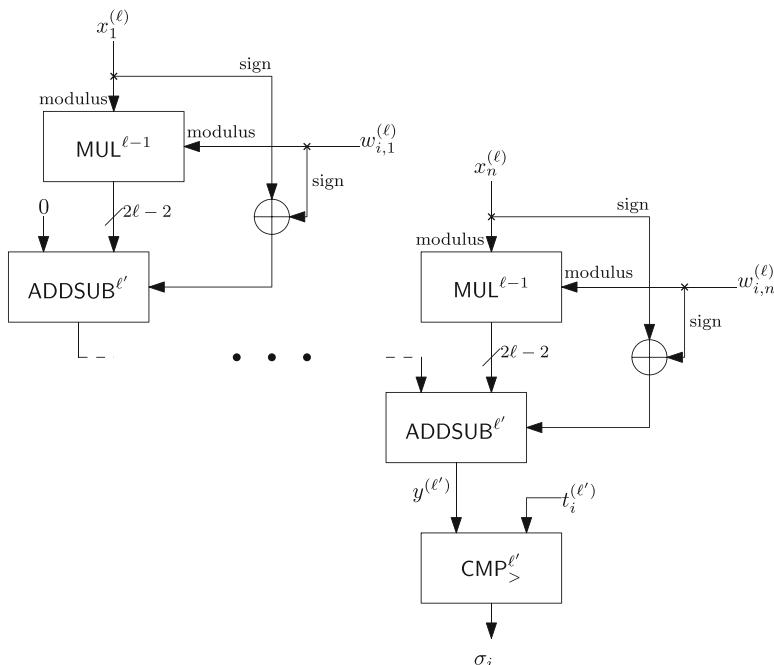


Fig. 22.4 Linear selection circuit (part of C) of a node

Table 22.4 Protocols for secure evaluation of private LBPs

Classification protocol	Moves	Asymptotic communication complexity		
		GC	OT	HE
Full-GC	2	$(2nd\ell^2 + nd\ell')3t$	$OT_t^{n\ell}$	
Hybrid	4	$6d\ell't$	$OT_t^{d\ell'}$	$(n + \lceil \frac{\ell'}{T-\tau} \rceil d)2T$

of the corresponding weight $w_{i,j}^{(\ell)}$ (by using the multiplication circuit $\text{MUL}^{\ell-1}$ [34]), while the sign of the product is determined by using XOR gates with the sign of the features and the corresponding weights as inputs. Afterwards, the outputs of the products are represented with $2\ell-2$ bits and each of them is summed or subtracted to the others, according to the corresponding product sign, by using $n-1$ $\text{ADDSUB}^{\ell'}$ circuits [34]. The scalar product result $y_i^{(\ell')} = \sum_{j=1}^n w_{i,j}^{(\ell)} \cdot x_j^{(\ell)}$ is then compared with the threshold $t_i^{(\ell')}$ by using a comparison circuit ($\text{CMP}_{>}^{\ell'}$) [34] that outputs σ_i . Finally the d values σ_i are input to a d -input gate that returns the string of the corresponding class. The resulting circuit is composed by $\approx 2nd\ell^2 + nd\ell'$ non-XOR 2-input gates and a d -input gate [4].

Considering only the precomputation of OT and transmitting the circuit during the online phase, the communication complexity of the full-GC protocol, shown in Table 22.4, corresponds to the complexity of the $OT_t^{n\ell}$ protocol (2t bits online for each input bit composing the feature vector) plus the size of the garbled circuit (3t bits for each non-XOR gates). The whole communication is performed in two rounds.

22.4.1.2 Hybrid Implementation

A secure LBP implementation can also be obtained by using a hybrid protocol, as shown in Fig. 22.5. The idea is to compute the products (the most expensive part in the full-GC solution) by relying on HE and then using GC to evaluate the comparison, whose HE implementation is expensive, and the classification tree.

We assume that \mathcal{C} has already generated a key-pair for the additively homomorphic encryption scheme and has sent the public key $pk_{\mathcal{C}}$ to \mathcal{S} . \mathcal{C} encrypts the elements of his composite feature vector and sends the encrypted values $\llbracket x_1^{(\ell)} \rrbracket, \dots, \llbracket x_n^{(\ell)} \rrbracket$ to \mathcal{S} . Using the additively homomorphic property, \mathcal{S} can compute the scalar product between these ciphertexts and the weights $\mathbf{w}_i^{(\ell)}$ as $\llbracket y_i^{(\ell')} \rrbracket = \llbracket \sum_{j=1}^n w_{i,j}^{(\ell)} x_j^{(\ell)} \rrbracket$, $1 \leq i \leq d$.

Afterwards, the values $y_i^{(\ell')}$ must be obliviously compared with the thresholds $t_i^{(\ell')}$. HE-based comparison is an expensive operation, hence a GC implementation is preferable. The ciphertexts $\llbracket y_i^{(\ell')} \rrbracket$ are converted into the corresponding garbled

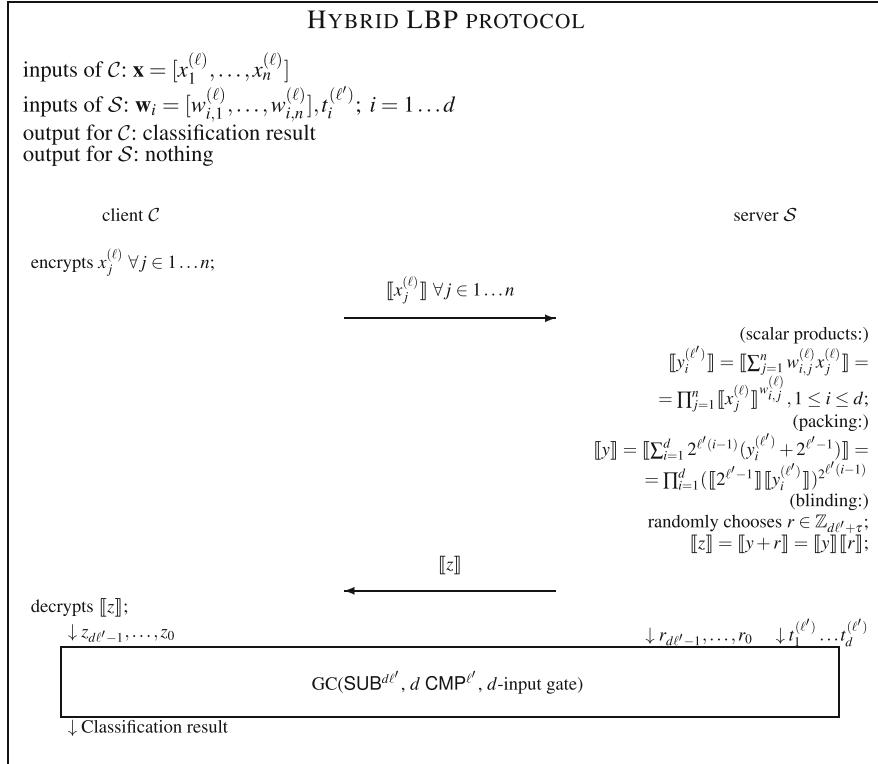


Fig. 22.5 Hybrid LBP protocol. For simplicity we assume that all the y_i values can be packed in a single ciphertext

values through a ‘‘HE to GC’’ interface (see Sect. 22.3.4). The $y_i^{(\ell')}$ values must be blinded before being sent to \mathcal{C} . To reduce the required bandwidth of the interface, several values can be packed together into a single ciphertext. We assume that the plaintext space of the HE cryptosystem (\mathbb{Z}_N) is large enough to contain the sum of the blinding value and the bitwise concatenation of d' values, i.e. $d' = \lfloor (T - \tau) / \ell' \rfloor$, where τ is a blinding security parameter. Hence $\lceil d/d' \rceil$ ciphertexts are obtained. In the following, we suppose for simplicity that only a ciphertext is obtained. To pack the values together, the encrypted values $-2^{\ell'-1} < y_i^{(\ell')} < 2^{\ell'-1}$ are shifted into the positive range $(0, 2^{\ell'})$ by adding $2^{\ell'-1}$. Afterwards, they are concatenated by computing $\llbracket y_p \rrbracket = \llbracket \sum_{i=1}^d 2^{\ell'(i-1)} (y_i^{(\ell')} + 2^{\ell'-1}) \rrbracket = \prod_{i=1}^d (\llbracket 2^{\ell'-1} \rrbracket \llbracket y_i^{(\ell')} \rrbracket)^{2^{\ell'(i-1)}}$. The packed ciphertext $\llbracket y_p \rrbracket$ now encrypts a $d\ell'$ bit value.

\mathcal{S} blinds the encrypted value $\llbracket y_p \rrbracket$ in order to hide the encrypted plaintext from \mathcal{C} . To do so, \mathcal{S} adds a randomly chosen value $r \in \mathbb{Z}_{d\ell'+\tau}$ under encryption before sending it to \mathcal{C} , who can decrypt the resulting ciphertext but does not get to know the plain value y_p . Afterwards, a circuit C' is evaluated by using GC. The circuit C' has the $d\ell'$ least significant bits of $y_p + r$ and r as inputs, together with the threshold

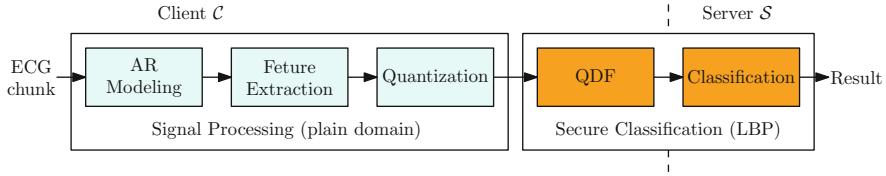


Fig. 22.6 Privacy-preserving ECG diagnosis

values $i_i^{(\ell')}$ and outputs the classification result. First of all the blinding value r is removed by using a $\text{SUB}^{d\ell'}$ block (a subtraction circuit [34] composed by $d\ell'$ non-XOR gates), then the result y_p is split into the d values $y_i^{(\ell')}$ that are compared with the threshold value $t_i^{(\ell')}$ in $d \text{ CMP}^{\ell'}$ blocks (ℓ' non-XOR gates each) and finally the decision tree is evaluated through a d -input gate.

The asymptotic communication complexity of the hybrid LBP protocol (shown in Table 22.4) consists of $n + \frac{d}{d'} \ell'$ Paillier ciphertexts of size $2T$ bits each, garbled circuits of size $\approx 2d\ell'$ non-XOR gates (each one having size $3t$), and a parallel OT protocol $OT_t^{d\ell'}$. Two rounds are necessary to send the homomorphic encryptions plus those of the underlying OT protocol (the garbled circuit can be sent with the last message of the OT protocol).

22.4.2 ECG Classification Through LBP and Quadratic Discriminant Functions

In order to describe how LBP's can be used for privacy-preserving ECG classifier, we observe that the classification algorithm based on the QDF functions and the classification tree (described in Sect. 22.2) is nothing but an LBP with an ℓ -bit representations composite feature vector $\mathbf{x}^{(\ell)}$ as attribute vector, and six nodes $P_i = \langle \mathbf{w}_i^{(\ell)}, 0 \rangle$, $i = 1, \dots, 6$, where $\mathbf{w}_i^{(\ell)}$ are ℓ -bit representations of the projection vectors. In this way, the general scheme for the privacy-preserving implementation of the classifier assumes the form given in Fig. 22.6.

The composite feature vector is computed by \mathcal{C} on plain data. Such a choice does not jeopardize the security of the system from the server's point of view, since \mathcal{S} is not interested in keeping the structure of the classifier secret, but only in preventing users from knowing the matrix $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_6]$ and the classification tree. On the contrary, all the steps from the projection onto the directions \mathbf{w}_i 's, until the final classification are carried out securely. Note that with respect to the overall architecture depicted in Fig. 22.1, a quantization block is introduced before the encryption of the composite feature vector. The need for such a block stems from the observation that the parameters a_1, a_2, a_3, a_4 resulting from the AR model estimation procedure are usually represented as floating point numbers, a representation that is not suitable for s.p.e.d. protocols which can be applied to

integer numbers only. For this reason, the elements of the composite feature vector \mathbf{x} and the coefficients of the matrix \mathbf{W} are quantized and represented in integer arithmetic for subsequent processing. Note that the choice of the quantization step, and consequently the number of bits used to represent the data (ℓ in the LBP terminology), is crucial since on one hand it determines the complexity of the overall secure protocol and on the other hand it has an impact on the accuracy of the ECG classification.

In contrast to the protocol described in Sect. 22.2, the sixth feature is omitted, obtaining the feature vector $\mathbf{f} = [a_1, a_2, a_3, a_4, n_1]$ and hence the composite feature vector \mathbf{x} is composed by 21 values. This omission is due to the fact that the removed feature can be expressed as $n_2 = 300 - n_1$, where 300 is the number of samples in each chunk, and in the QDFs each product involving n_2 (its square values and its products with other features) can be expressed as a function of n_1 .

22.4.2.1 Quantization Error Analysis

The quantization error introduced passing from \mathbf{x} to $\mathbf{x}^{(\ell)}$ and from \mathbf{w}_i to $\mathbf{w}_i^{(\ell)}$ impacts on the classification accuracy. Such impact is analyzed to determine the number of bits needed to represent the attribute vector and the linear combination vectors of the LBP. The value of ℓ influences the complexity of the secure classification protocol.

To start with, we observe that quantization is applied to the composite feature vector \mathbf{x} , that is used to compute the vector \mathbf{y} , through multiplication with the matrix \mathbf{W} . After such a step, only the signs of \mathbf{y} are retained, hence it is sufficient to analyze the effect of quantization until the computation of the sign of \mathbf{y} . As to the processing steps carried out by the client prior to quantization, we assume that all the blocks until QDF are carried out by using a standard double precision floating point arithmetic. In order to simplify the notation, we consider the computation of one coefficient of the vector \mathbf{y} . The function to be computed is a simple inner product: $\mathbf{y} = \mathbf{w}\mathbf{x} = \sum_j w_j x_j$ where the index i has been omitted, and w_j and x_j are real numbers. The quantized version of the above variables can be expressed as follows:

$$\begin{aligned} w_{q,j} &= \lfloor k_1 w_j \rfloor, \\ x_{q,j} &= \lfloor k_2 x_j \rfloor, \end{aligned} \tag{22.5}$$

where k_1 and k_2 are positive integers chosen so that the values involved in the computation do not exceed a given bitlength and the error is lower than a target. From a worst case analysis (see the details in [7]), it results that 54 bits are necessary to represent each of the 21 features in \mathbf{x} and the elements in \mathbf{W} and guarantee a final error lower than 10^{-5} . However, we may expect that in practice a good classification accuracy can be obtained also with less than 54 bits. To investigate this aspect, a simulator has been implemented to detect experimentally which is the minimum number of bits needed. The results obtained by running the simulator on the ECG dataset are shown in Fig. 22.7. This figure shows that $\ell = 44$ is sufficient to guarantee the same performance of a non-quantized implementation.

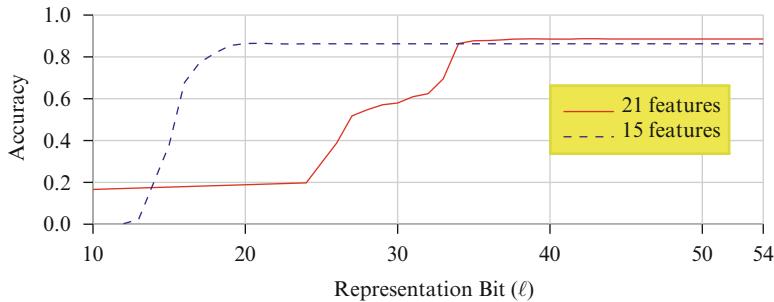


Fig. 22.7 Classification accuracy of dataset using 21 and 15 features

While classification accuracy is a primary goal in plain domain implementations, where the protocol runs in negligible time, in s.p.e.d. applications the protocol complexity is so high that a small accuracy loss can be accepted if it results in a significant complexity improvement. In order to further speed up the protocol, a version of the ECG classifier with a reduced number of features has been tested. Specifically, the composite feature vector size is reduced by eliminating also the feature n_1 . In this way, we obtain a 15-coefficient \mathbf{x} . Obviously the reduction of the feature space also results in a reduction of the accuracy, but this reduction is quite negligible: experiments, in fact, indicate that the accuracy decreases only from 88.33 % to 86.30 %. On the other hand, as it will be shown later, by removing one feature we gain a lot from a complexity point of view. Such a gain is already visible in Fig. 22.7, showing that with the reduced set of features, a value of ℓ as low as 24 is enough to obtain the same performance of a non-quantized version of the classifier.

22.4.3 ECG Classification Through LBP and Linear Discriminant Functions

In the previous section, we have shown that the complexity of ECG classification can be reduced by discarding a single feature at the expense of a small accuracy loss. In this section we show that a greater complexity reduction can be obtained by using Linear Discriminant Functions (LDF) instead of QDF, again with a loss of few percentage points in the accuracy of the classification.

In the classification tree defined in Sect. 22.2, linear functions are used instead of quadratic functions. Hence the feature vector $\mathbf{f} = [a_1, a_2, a_3, a_4, n_1]$ is used, i.e. $\mathbf{x} = \mathbf{f}$. This provides many advantages. First of all also, the bitlengths of the feature vector and weight vectors \mathbf{w}_i are reduced and the number of products necessary to compute $\mathbf{y} = \mathbf{Wx}$ is reduced as well. Moreover, the maximum value that a single feature can assume is also reduced, since no second order elements are present. In fact, from the worst case analysis we obtain that 33 bits are sufficient to correctly

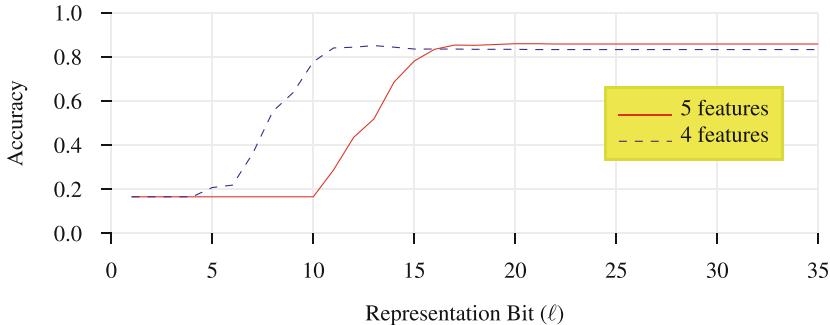


Fig. 22.8 Classification accuracy of dataset using 5 and 4 features

represent the magnitude of features and weights. Indeed, as in the QDF case, less bits could be used to represent the values. Once again the minimum number of bits can be determined by measuring the accuracy as a function of the bitlength used to represent the inputs. Figure 22.8 shows the results obtained with $n = 5$ and $n = 4$ features respectively. We can see that an accuracy of 85.95 % can be reached by using $\ell = 22$ (excluding sign) when the feature vector is composed by 5 features. By discarding n_1 the accuracy decreases to 83.33 %. However, as Fig. 22.8 shows, a similar accuracy can be guaranteed already with $\ell = 15$ bits, where errors respect the floating point implementation make the classification accuracy increase to 83.35 %.

22.4.4 Complexity Analysis

The communication complexity of the Hybrid and the GC protocols is estimated by relying on the analysis carried out in the previous section. The communication complexity is evaluated by considering a C++ implementation of the two protocols. The security parameters are the bitlength T of the RSA modulus for Paillier encryption [39], and the bitlength t of the symmetric security parameter for the GC protocol. For the tests, the following security parameters are chosen according to common recommendations [23]: $T = 1248$, $t = 80$ for short-term security (recommended use up to 2010³), $T = 2432$, $t = 112$ for medium-term security (up to 2030) and $T = 3248$, $t = 128$ for long-term security (up to 2050).

The following configurations have been considered:

- φ1: QDF, $n = 21$ and $\ell = 54$ bits, as obtained from the theoretical estimations;
- φ2: QDF, $n = 21$ and $\ell = 44$ bits, the lower value obtained from the practical tests;

³Tests were performed in 2009. Anyway these values are still widely used today.

Table 22.5 Estimated communication complexity of LBP with QDF

Configuration	Features	n	ℓ	Accuracy (%)	Protocol type	Short-term security	Medium-term security	Long-term security
ϕ_1	5	21	54	88.33	GC hybrid	21.5 MB 65.8 KB	30.1 MB 95.6 KB	34.4 MB 111.6 KB
ϕ_2	5	21	44	88.33	GC hybrid	14.4 MB 55.2 KB	20.1 MB 80.8 KB	23.0 MB 94.7 KB
ϕ_3	4	15	24	86.30	GC hybrid	3.1 MB 31.7 KB	4.4 MB 46.9 KB	5.0 MB 55.3 KB
ϕ_4	5	5	22	85.95	GC hybrid	0.9 MB 26.0 KB	1.2 MB 37.2 KB	1.4 MB 43.1 KB
ϕ_5	4	4	21	83.33	GC hybrid	0.6 MB 24.1 KB	0.9 MB 34.4 KB	1.0 MB 39.8 KB
ϕ_6	4	4	15	83.35	GC hybrid	0.3 MB 17.8 KB	0.5 MB 25.5 KB	0.5 MB 29.6 KB

- ϕ_3 : QDF, $n = 15$ and $\ell = 24$ bits, obtained by discarding feature n_1 ;
 ϕ_4 : LDF, $n = 5$ and $\ell = 22$ bits, the smallest bitlength guaranteeing an acceptable accuracy;
 ϕ_5 : LDF, $n = 4$ and $\ell = 21$ bits, n_1 is discarded;
 ϕ_6 : LDF, $n = 4$ and $\ell = 15$ bits, the value bitlength is decreased respect to ϕ_5 .

The estimated communication complexity of the ECG classification protocols is provided in Table 22.5. Offline precomputation of OT is used, but not the transmission of the circuit during precomputation. Note that since $\ell' = 112$ (in the worst case of ϕ_1) and $d = 6$, all the $y_i^{(\ell')}$ values can be packed in a single ciphertext.

The computation complexity of the first three configurations has been measured by running the hybrid and GC protocols, both implemented in C++ using the Miracl library.⁴ The measured complexities for short-term security are shown in Table 22.6, as reported in the original paper [5]. The third configuration has been tested even for medium-term security. The tests were performed on two PCs with 3 GHz Intel Core Duo processor and 4 GB memory connected via Gigabit Ethernet. The computation complexity for the client and the server are separated into CPU time and total time which additionally includes data transfer and idle times. We highlight that in the tests, OT has not been precomputed and garbled row reduction was not implemented. From these measurements we draw the following conclusions.

Different strategies can be used to decrease the protocol complexity. By discarding only the sixth feature in the QDF and quantizing the features with 24 bits (§3), we have a slight loss of accuracy of the protocol ($\sim 2\%$). A similar accuracy loss ($\sim 2.4\%$) has been obtained by opting for a LDF (ϕ_4) instead of a QDF. This is

⁴<http://www.shamus.ie>.

Table 22.6 Performance of protocols for secure ECG classification through LBP

Configuration	Features	n	ℓ	Protocol type	Computation			
					Client (s)		Server (s)	
					CPU	Total	CPU	Total
ϕ_1	5	21	54	Hybrid	2.3	35.4	5.4	34.2
				GC	7.2	64.5	17.3	64.7
ϕ_2	5	21	44	Hybrid	2.0	29.0	4.8	27.6
				GC	4.7	48.5	11.5	48.8
ϕ_3	4	15	24	Hybrid	1.3	18.7	3.3	16.2
				GC	1.3	17.5	3.1	19.2
ϕ_3^a	4	15	24	Hybrid	6.5	40.5	16.3	30.9
				GC	3.0	20.4	4.6	20.8

^a medium-term security

acceptable because an ECG may be processed in a very short time, opening the way to real time processing. On the other hand, accuracy decreases by other 2.5 % points in configurations ϕ_5 and ϕ_6 .

The performance improvements obtained in the configuration ϕ_{2-6} are due to a reduced number of features and smaller values of ℓ . The first one results in a lower number of products, affecting particularly the complexity of the GC implementation, and the bandwidth of the hybrid implementation; the second one affects the number of gates composing each block of the circuits, especially products.

The bandwidth in the GC protocol (MB) is an order of magnitude larger than that of the hybrid protocol (KB). However, the GC protocol requires only two transmission rounds, while the hybrid protocol requires 4 rounds. Configuration ϕ_4 is indeed preferable to the ϕ_{1-3} configurations making use of QDF, but the accuracy loss (3 %) from configuration ϕ_4 to configuration ϕ_5 is not acceptable given the slight performance improvement. Only the small complexity of configuration ϕ_6 can justify the choice of using both LDF and four features. On the other hand the privacy preserving implementation does not provide the same results of a floating implementation in the plain domain.

The computation complexity has been measured as runtimes for configurations ϕ_{1-3} . In ϕ_1 and ϕ_2 the runtime of the Hybrid protocol is three times lower than that of the GC protocol (CPU time) and two times faster if we consider the total time, whereas for the optimized test case ϕ_3 both protocols have approximately the same computation complexity. Configurations ϕ_{4-6} has not been tested. However, given the smaller parameter sizes with respect to configuration ϕ_3 , we expect that configurations ϕ_{4-6} have runtimes in the order of some seconds.

For medium-term security, the GC protocol is substantially better than the Hybrid protocol. Increasing the security parameters has a more dramatic effect on the computational complexity of the Hybrid protocol than on that of the GC protocol

(see test $\phi 3$ vs. $\phi 3^a$). This effect results from the asymmetric security parameter T being almost doubled, whereas the symmetric security parameter t is only slightly increased.

We can derive that by representing the quantized features with a lower number of bits (the minimum that permits a correct evaluation) and omitting the feature related to the AR model error or opting for the LDF, the communication and computation complexities decrease significantly, thus making the small accuracy loss acceptable. However the highest accuracy loss due to the joint use of LDF and 4 features is not justified by the performance gain.

22.4.4.1 More Efficient LBP Implementations

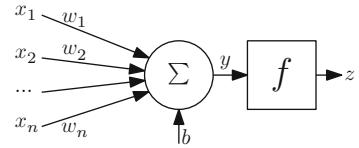
We conclude this section underlying that a more efficient LBP implementation can be obtained by introducing precomputation in the full-GC and Hybrid LBP protocols, i.e. not only by using OT precomputation, but also moving circuit garbling and transmission to an offline phase (details in [7]).

22.5 Privacy Preserving Classification by Using Neural Network

In this section, we present an alternative privacy preserving ECG classification protocol based on artificial Neural Networks (NN) [7]. NNs are well-known machine learning structures used in many different fields ranging from approximation to classification. NNs are widely used as classifiers and, in general, they give good results if the set used to train the network is representative of all the considered classes and the generalization grade is good enough (see [28] or [2]).

22.5.1 Neural Network Design

Finding the right topology for a NN is not a simple task due to the fact that NNs have several degrees of freedom including: number of hidden layers, neurons per hidden layer and form of activation functions. In most cases, a two layer NN is sufficient to obtain good classification accuracy, therefore in the rest of the chapter we focus on NNs with two layers, that is NNs in which the inputs are connected to a hidden layer, that, in turn, is connected to the output layer. Each neuron weights the inputs through a scalar product and uses the result in an activation function. As shown in Fig. 22.9, each neuron, or *perceptron*, evaluates a scalar product among the weights $\mathbf{w} = (w_1, w_2, \dots, w_n)$ and the input vector $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ and adds up the bias b :

Fig. 22.9 A perceptron

$$y = b + \mathbf{w} \cdot \mathbf{x} = b + \sum_{i=1}^n w_i x_i, \quad (22.6)$$

after this, the activation function is applied producing the neuron output: $z = f(y) = f(b + \mathbf{w} \cdot \mathbf{x})$.

A NN is composed by a cascade of layers. Usually only two layers are used. The first layer is named the hidden layer and is composed by n_h neurons connected to the inputs, while the second one is the output layer, composed by n_o neurons. \mathbf{W}_h is a matrix of size $n_h \times n$ whose elements are the weights of the connections between the inputs and the hidden layer, while \mathbf{b}_h is a vector of length n_h that contains the biases of the neurons. The element $\mathbf{W}_h(i,j) = w_{h;i,j}$ is the coefficient used to weight the j th input in the i th node of the hidden layer. The output of the hidden layer, a vector of length n_h , is

$$\mathbf{z}_h = f_h(\mathbf{y}_h) = f(\mathbf{b}_h + \mathbf{W}_h \mathbf{x}), \quad (22.7)$$

where the i th element $z_{h;i} = f_h(y_{h;i}) = f_h(b_{h;i} + \sum_{j=1}^n w_{h;i,j} x_j)$ is the output of the corresponding neuron, and the activation function $f_h(\cdot)$ is applied component-wise to all the values of the input vector. The output of the hidden layer is then connected to the neurons of the output layer. The weights of the output layer are arranged in a matrix \mathbf{W}_o of size $n_o \times n_h$ whose element $\mathbf{W}_o(i,j) = w_{o;i,j}$ is the coefficient used to weight the output of the j th node of the hidden layer in the i th node of the output layer. The bias vector of the output layer is \mathbf{b}_o and has length n_o , while the output of the neuron is

$$\mathbf{z}_o = f_o(\mathbf{y}_o) = f_o(\mathbf{b}_o + \mathbf{W}_o \mathbf{z}_h) \quad (22.8)$$

where $z_{o;k} = f_o(y_o) = b_{o;k} + \sum_{i=1}^{n_h} w_{o;k,i} z_{h;i}$ is the output of the k th neuron of the output layer.

The activation function f_h in the hidden layer can differ from the activation function in the output layer. In fact in the NN described below the neurons of the hidden layer apply the `satlin`(\cdot) function (details below) while the output layer applies the identity function, i.e. $\mathbf{z}_o = \mathbf{y}_o$. This choice is motivated by the necessity of solving conflicts that may arise. In fact we expect that only one output neuron returns a positive value, but sometimes two or more output neurons erroneously classify the input in the associated classes. If more neurons have a positive output, the one returning the highest value is chosen for classification through a maximum selection tree, hence the NN classification result is $o = \text{ARGMAX}\{\mathbf{z}_o\}$.

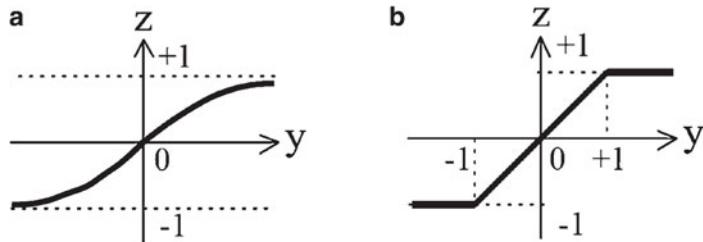


Fig. 22.10 Transfer functions. (a) tansig. (b) satlin

We now describe the NN design for the ECG classification. The number of inputs is determined by the number of features the classifier relies on. Only the features coming from the AR model are used and hence $n = 4$. The number of output layers corresponds to the number of diseases the NN should distinguish, so we have $n_o = 6$. The degrees of freedom we have, then, are n_h , the number of neurons in the hidden layer, and the activation function in the hidden layer.

The activation function is chosen so to ease a s.p.e.d. implementation. A common choice would be the Hyperbolic Tangent Sigmoid transfer function $\text{tansig}(x) = 2/(1 + e^{-2x}) - 1$, widely used in NNs (Fig. 22.10a). Unluckily, the $\text{tansig}(\cdot)$ function is difficult to implement in a s.p.e.d. protocol, hence the $\text{satlin}(\cdot)$ transfer function (Fig. 22.10a), defined by

$$f(y) = \text{satlin}(y) = \begin{cases} 1 & \text{if } y > 1 \\ y & \text{if } -1 < y < 1 \\ -1 & \text{if } y < -1, \end{cases} \quad (22.9)$$

is used.

To choose n_h , some tests have been carried out trying to reach the same accuracy provided by the LBP-based classifier in configuration $\phi 3$. There is not a rule to choose the correct number of neurons for the hidden layer. NNs with a number of neurons varying from 4 to 15 neurons have been tested. A training set was built by using as samples the pair: (\mathbf{x}, \mathbf{o}) where, as said before, \mathbf{x} is the feature vector (a_1, a_2, a_3, a_4) and \mathbf{o} is a six-component vector, having value 1 for the index of the class the ECG signal belongs to and -1 elsewhere.⁵ In the experiments the dataset has been split into a training set (containing 140 ECG sequences) and a test set (with the remaining 60 signals).⁶ As shown in Fig. 22.11, the smallest NN giving a

⁵This kind of Neural Network is often called NN with *fired* output.

⁶NN network training requires a large dataset, hence authors had the necessity to choose a bigger training set than the one used for the LBP implementation, where the training and test dataset size have been chosen according to [21].

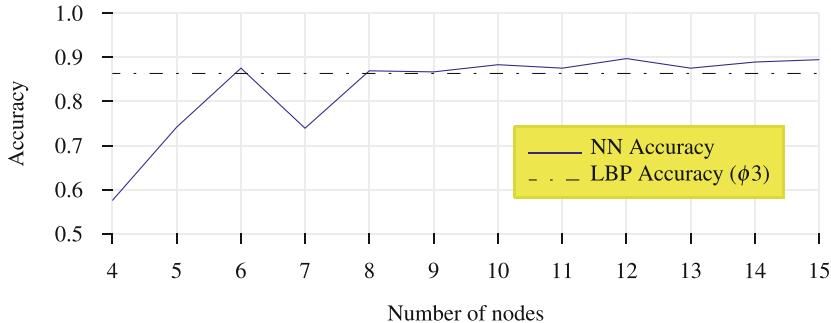


Fig. 22.11 Classification accuracy as a function of the number of nodes in the hidden layer and satlin as activation function

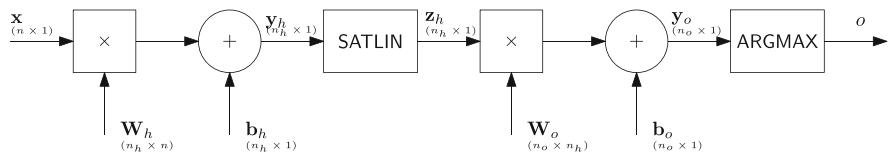


Fig. 22.12 Neural network structure. In the ECG classification protocol $n = 4$, $n_h = 6$ and $n_o = 6$

classification accuracy larger than 86.30 % has $n_h = 6$ neurons in the hidden layer. The steps composing the NN are shown in Fig. 22.12.⁷

22.5.2 Quantized Neural Network Classifier

For a s.p.e.d. implementation we need to introduce a quantized version of the NN classifier described so far. Specifically, the multipliers q_i , q_h , q_o are introduced to quantize, respectively, the inputs of the NN, and the parameters (weights and biases) of the hidden and the output layers. In the following, we use the symbols ℓ^i , ℓ^h , ℓ^o to indicate, respectively, the number of bits needed to represent the quantized version of the inputs and the quantized parameters of the hidden and output layers, including the sign bits.

By defining the quantized input as $\mathbf{x}_q = \lfloor q_i \mathbf{x} \rceil$ and the quantized weights as $\mathbf{W}_{hq} = \lfloor q_h \mathbf{W}_h \rceil$, the quantized output vector \mathbf{z}_{hq} of the hidden layer is:

$$\mathbf{z}_{hq} = \text{QSATLIN}(\mathbf{y}_{hq}) = \text{QSATLIN}(\mathbf{b}_{hq} + \mathbf{W}_{hq} \mathbf{x}_q), \quad (22.10)$$

⁷Many other NNs have been trained by changing the number of neurons and the activation functions, as well as the training method, before choosing the NN described here. See [7] for details.

where the biases $\mathbf{b}_{hq} = \lfloor q_i q_h \mathbf{b}_h \rfloor$ have been multiplied by both q_i and q_h , to make the bias homogeneous with the term $q_i q_h \mathbf{W}_h \mathbf{x}$. They need to be represented with $\ell^{bh} = \ell^i + \ell^h - 1$ bits. The components of \mathbf{y}_{hq} are obtained as the sum of 5 components: the products between the inputs and the weights and the bias. Considering that the magnitude of the components of \mathbf{x}_q and \mathbf{W}_{hq} are represented with $\ell^i - 1$ and $\ell^h - 1$ bits respectively, each element in \mathbf{y}_{hq} needs $\ell^{yh} = \ell^i + \ell^h + 2$ bits, including the sign. The SATLIN function of Eq. (22.9) is replaced with its quantized version defined as follows:

$$\text{QSATLIN}(y_{hq}) = \begin{cases} q_i q_h & \text{if } y_{hq} \geq q_i q_h \\ y_{hq} & \text{if } -q_i q_h < y_{hq} < q_i q_h \\ -q_i q_h & \text{if } y_{hq} \leq -q_i q_h \end{cases} \quad (22.11)$$

where saturation occurs when the magnitude of the input is equal to $q_i q_h$ corresponding to a unitary magnitude of the non-quantized inputs. Due to saturation, the output of QSATLIN can be represented with less bits than its input, namely $\ell^q = 1 + \lceil \log_2(q_i q_h) \rceil$ bit at most (one bit represents the sign), so each component in \mathbf{z}_{hq} requires at most ℓ^q bits, less than its input \mathbf{y}_{hq} .

A similar analysis can be applied to the output layer, where the $\mathbf{b}_{oq} = \lfloor q_o q_h q_i \mathbf{b}_o \rfloor$ components are represented with $\ell^{bo} = \ell^q + \ell^o - 1$ bits and

$$\mathbf{z}_{oq} = \mathbf{y}_{oq} = \mathbf{b}_{oq} + \mathbf{W}_{oq} \mathbf{y}_{hq}, \quad (22.12)$$

where each component of \mathbf{y}_{oq} requires $\ell^{yo} = \ell^q + \ell^o + 2$ bits (as before one bit is used for the sign). At this point, the output of the maximum function that completes the classification is $o = \text{ARGMAX}\{\mathbf{z}_{oq}\}$ and, since o is just the index of the largest component, the number of bits necessary to represent it is the logarithm of the number of neurons in the output layer, i.e. $\lceil \log_2 n_o \rceil = 3$ bits. Finally the NN classification result is $o = \text{ARGMAX}\{\mathbf{y}_o\}$ that can be represented with $\lceil \log_2 n_o \rceil = 3$ bits.

22.5.2.1 Representation vs. Classification Accuracy

We are now ready to determine the minimum number of bits necessary to represent the values involved in the computations, in order to obtain the same accuracy of a floating point implementation. This is a crucial step, since the size in bits of the input features and that of the classifier parameters have an immediate impact on the complexity of the s.p.e.d. implementation.

The minimum number of bits necessary to reach the same classification accuracy of the LBP classifier has been obtained by running a simulator that evaluates the classification accuracy of the NN in the case $\ell^i = \ell^h = \ell^o$, obtaining the classification accuracy shown in Fig. 22.13. To guarantee the same classification obtained by the LBP classifier, at least $\ell^i = \ell^h = \ell^o = 13$ bits must be used. These

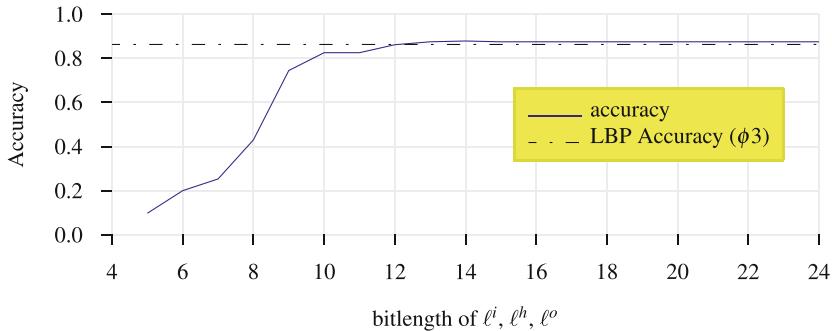


Fig. 22.13 Classification accuracy as a function of ℓ^i, ℓ^h, ℓ^o

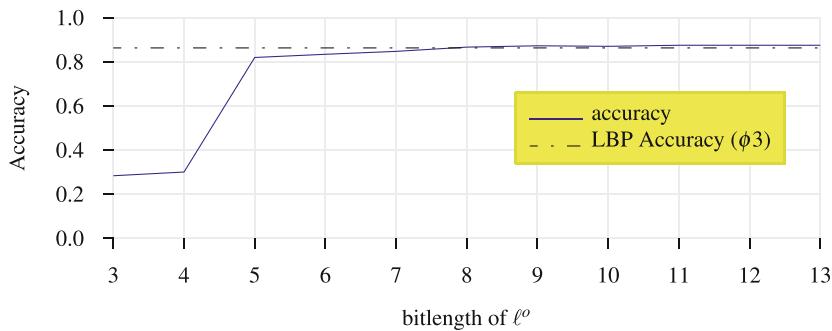


Fig. 22.14 Classification accuracy in function of ℓ^o , with $\ell^i = \ell^h = 13$

numbers of bits (namely ℓ^i and ℓ^h) are those necessary to correctly represent the integer part of the input features and NN parameters multiplied by the quantization factors q_i and q_h .

The results of the scalar products in the hidden layer are used as input of the QSATLIN(\cdot) function that gives an output with magnitude bounded by $q_i q_h$, hence it can be represented with $\ell^q = \lceil \log_2 q_i q_h \rceil + 1 = 18$ bits, including the sign bit. As a further test, the number of bits used for the output layer has been reduced once the bitlength of the parameters of the hidden layer has been fixed. As shown in Fig. 22.14, letting $\ell^o = 8$ is sufficient to guarantee the same accuracy.

22.5.3 Privacy-Preserving GC-Based NN Classifier

In this section we describe a GC implementation of the NN (this section is based on the system presented in [7]).

The inputs to the NN are determined according to the analysis carried out in the previous section and shown in Table 22.7. Secrets are associated to \mathcal{C} inputs by

Table 22.7 Inputs to the GC implementation of the neural network

Input	Owner	Bitlength
\mathbf{x}_q	\mathcal{C}	$n \cdot \ell^i = 4 \cdot 13 = 52$
\mathbf{W}_{hq}	\mathcal{S}	$n_h \cdot n \cdot \ell^h = 312$
\mathbf{b}_{hq}	\mathcal{S}	$n_h \cdot \ell^{bh} = 150$
\mathbf{W}_{oq}	\mathcal{S}	$n_o \cdot n_h \cdot \ell^o = 288$
\mathbf{b}_{oq}	\mathcal{S}	$n_h \cdot \ell^{bo} = 150$

evaluating an OT_t^{52} , while \mathcal{S} directly transmits the secrets associated to his 900 bits. The final output, represented in 3 bits, is provided to \mathcal{C} , while all the intermediate values are hidden to both \mathcal{C} and \mathcal{S} . The circuit is constructed by instantiating each block of Fig. 22.12 through boolean circuits, as described below.

Hidden layer—scalar products: The inputs \mathbf{x}_q and \mathbf{W}_{hq} are given in sign-magnitude representation so that their component-wise product can be easily computed: the magnitude is the product of the input magnitudes using $n \cdot n_h$ MUL blocks [34] while the sign is computed “for free” by XORing the input signs. Now, depending on the sign, the magnitudes of the products are added to or subtracted from \mathbf{b}_{hq} by using a cascade of n_h ADDSUB blocks [34]. The output \mathbf{y}_{hq} is a vector with $n_h = 6$ components of 28-bit signed values in 2’s complement representation.

Hidden layer—activation functions: Afterwards, for each of the $n_h = 6$ neurons the QSATLIN activation function is evaluated: first, the 28-bit input is converted from 2’s complement into sign/magnitude representation with an ADDSUB block. Then, the QSATLIN function, defined in Eq. (22.11) is evaluated by a circuit implementing $QSATLIN(y_{hq}) = \text{sign}(y_{hq}) \cdot \min(\text{abs}(y_{hq}), q_i q_h)$ (we omit the index of the neuron for simplicity). The minimum is computed by comparing the magnitude of y_{hq} with $q_i q_h$ using a CMP block [34]. Depending on the outcome of this comparison, the magnitude of the outcome is either the magnitude of y_{hq} or $q_i q_h$ selected with a MUX block (a multiplexer circuit [34]). The output \mathbf{z}_{hq} is a vector of $n_h = 6$ components of $\ell^q = 18$ -bit signed values in sign/magnitude representation.

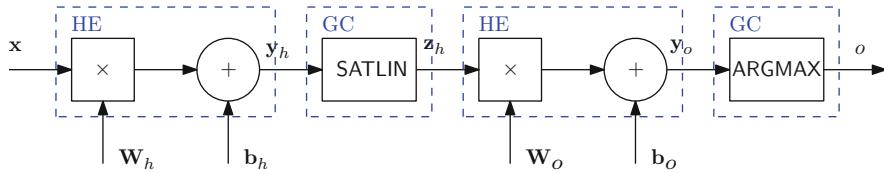
Output layer—scalar products: The value $\mathbf{y}_{oq} = \mathbf{b}_{oq} + \mathbf{W}_{oq} \mathbf{y}_{hq}$ is computed similarly to $\mathbf{b}_{hq} + \mathbf{W}_{hq} \mathbf{x}_q$. The output \mathbf{y}_{oq} is a vector with $n_o = 6$ components of 28-bit signed values in 2’s complement representation.

Output layer—maximum selection: Finally, the index of the maximum value \mathbf{o} (3 bits) is provided to \mathcal{C} and is determined with an ARGMAX block that selects the maximum input value and returns his identifier (the index of the node) [34].

In total, the circuit has 52 input wires corresponding to the inputs from \mathcal{C} , 900 input wires for the inputs belonging to \mathcal{S} , and at most 16,921 2-input non-XOR gates. The costs for this NN-based ECG classification protocol with symmetric parameter t are summarized in Table 22.8.

Table 22.8 Complexity of NN-based ECG classification protocol

Implementation	Moves	Short-term security	Medium-term security	Long-term security
Full-GC	2	505.5 KB	707.8 KB	808.9 KB
Hybrid	8	61.6 KB	98.8 KB	120.3 KB

**Fig. 22.15** Hybrid implementation of the neural network

22.5.4 Privacy-Preserving Hybrid NN Classifier

The NN can be also implemented by using a hybrid protocol, alternating HE and GC subprotocols, as shown in Fig. 22.15. The following analysis assumes that the same setting (and bit-lengths) used so far are adopted. As before, \mathcal{C} inputs the encrypted vector $[\![\mathbf{x}_q]\!]$, while \mathcal{S} inputs \mathbf{W}_{hq} , \mathbf{b}_{hq} , \mathbf{W}_{oq} and \mathbf{b}_{oq} . The final output is provided to \mathcal{C} , while all the intermediate values are hidden to both \mathcal{C} and \mathcal{S} .

Hidden layer—scalar products: The scalar products in the first step are implemented by using HE. \mathcal{C} sends the encrypted features to \mathcal{S} which evaluates six scalar products on encrypted data. Then \mathcal{S} packs the resulting \mathbf{y}_{hq} in a single ciphertext and blinds it with a random value R_h . Finally the value is sent back to \mathcal{C} . Considering that the scalar product results can be represented by using 28 bits, the packed value is composed by 168 bits and the random value has to be represented with $168 + \tau$ bits (usually $\tau = t$). In total 5 ciphertexts are transmitted in 2 moves for a total bandwidth of $10T$ bits.

Hidden layer—activation functions: The QSATLIN activation functions are evaluated by using GC. \mathcal{C} decrypts the ciphertext received and inputs the 168 least significant bits to a GC, through an online OT_t^{168} , while \mathcal{S} inputs the 168 least significant bits of R_h and other 6 random values, each one represented with $18 + \tau$ bits, used to blind the outputs of the activation functions. The circuit is composed by the SUB block that removes the obfuscation, the QSATLIN circuit (as described in Sect. 22.5.3) and finally 6 ADD circuits needed to blind the activation function outputs \mathbf{z}_{hq} . This step requires 2 moves and the transmission of $168 \cdot t$ bits for the OT, $168 \cdot t + 6 \cdot (18 + \tau) \cdot t$ bits as \mathcal{S} 's input secrets and $540 + 6\tau$ non-XOR gates of size $3t$.

Output layer—scalar products: \mathcal{C} encrypts the 6 blinded outputs obtained by the GC section and transmits the ciphertexts to \mathcal{S} , which removes the blinding and then evaluates 6 scalar products on encrypted data. Then \mathcal{S} packs the result in a single ciphertext and blinds it with a random value R_o . Finally the value is

sent back to \mathcal{C} . Considering that the scalar product results can be represented by using 28 bits, a single pack of 168 bits is obtained and the random value must be represented with $168 + \tau$ bits. In total 7 ciphertexts are transmitted in two rounds for a total bandwidth of $14T$ bits.

Output layer—maximum selection: In the last step, the index of the maximum value is retrieved by using GC. \mathcal{C} decrypts the received ciphertext and inputs the 168 least significant bits into a GC, through 168 online OT's, while \mathcal{S} inputs the 168 least significant bits of R_o . The circuit is composed by the SUB block that remove the obfuscation and the ARGMAX block [34]. In this phase, requiring 2 moves, $168 \cdot 2t$ bits are transmitted for the OT, $168 \cdot t$ bits are transmitted as \mathcal{S} 's input secrets and 427 non-XOR gates of size $3t$ are sent from \mathcal{S} to \mathcal{C} .

The communication complexity of the hybrid protocol is shown in Table 22.8.

22.5.5 Comparison with the LBP Solution

Comparing Tables 22.8 and 22.5, we can see that the full-GC based NN requires a lower bandwidth than LBP (505.5 KB for NN and 3.1 MB for LBP with short term security). The bandwidth of the hybrid NN decreases with respect to the full-GC NN, but is higher than that of the hybrid LBP, making the latter more efficient. The reason is that, as shown in Sect. 22.3, the 12 bit values to be multiplied are so small that the HE and GC implementations are roughly equivalent. The hybrid implementation of the NN requires the use of homomorphic protocols for the scalar products in both layers, the first one followed by a circuit implementing the QSATLIN function, and the second one followed by the ARGMIN circuit. The small benefit given by the use of HE is canceled out by the need to implement three interfaces between GC and HE subprotocols. The eight moves of the hybrid NN are also relevant, as each move introduces some delay due to the latency of the network.

On the other hand, circuit garbling and transmission can be moved on a setup phase, leaving only OT and circuit evaluation in the online phase, while few basic operations (for example the computation of exponentiations involving only random values) can be precomputed in the HE part.

In summary we can state that the LBP classifier is preferable from a communication complexity point of view when considering the total amount of data sent in setup and the online phases. On the other hand, the NN protocol relies only on fast symmetric encryption operations, hence resulting in a better performance from a computational complexity perspective, an advantage that becomes more significant for long term security, since the security parameters of asymmetric cryptosystems are going to increase more rapidly than those of symmetric cryptosystems.

By considering the classifier structures underlying the two protocols, we see that the NN ensures a twofold advantage since: (1) it allows to work on a smaller feature vector (four features instead of the 15 components of the composite feature vector required by the LBP classifier), and (2) it requires a smaller number of bits for

the representation of the feature vector and the parameters of the classifier. This is partially due to the presence of hard limiting activation functions avoiding that the inner results of the computation grows in magnitude. Due to the above properties the GC implementation of the NN protocol does imply a too large penalty for the necessity of working entirely with boolean instead of arithmetic circuits.

With respect to LBP with LDF, NN has a slightly higher complexity, but it guarantees a better accuracy.

We conclude our discussion by observing that the complexity of both protocols depends on the number of features used to classify the ECG signals. In the NN case, the dependence of the size of the classifier on the number of features is not easy to determine. On one hand, it results in an increase of the size of the input layer of the NN, with a linear impact on the complexity of the part of the protocol corresponding to the computation of the input of the hidden layer. On the other hand, it is likely that the number of neurons in the hidden layer will have to increase as well thus resulting in a superlinear dependence of the complexity on the number of features. The overall complexity increases, however, the increase is likely to be less than quadratic, given that the size of the output layer will remain constant. In the LBP case the dependence is at least quadratic due to the inclusion within the composite feature vector of the quadratic terms.⁸ For this reason, we expect that the NN structure is going to become even more advantageous if the number of features considered by the classifier increases.

22.6 Privacy Preserving Quality Evaluation

In order to avoid that possible errors due the inexperience of patients, who have not experience with placing the ECG electrodes, properly impair the result of the classification, it is necessary that the quality of the ECG signal is evaluated by the server before running the classifier. For the above reason, in this section we describe a protocol which allows the service provider to evaluate the quality of the to-be-classified ECG signal, without disclosing any information about the signal itself. The protocol first computes the SNR (Signal to Noise Ratio) of ECG signal, then it takes a binary decision about the quality, to decide whether to go on with the classification or not.

22.6.1 SNR Evaluation in the Encrypted Domain

The quality measure we are looking for belongs to the class of no-reference measures, since, in the scenario examined in this chapter, a noise-free version of

⁸Of course the size of the decision tree may change as well.

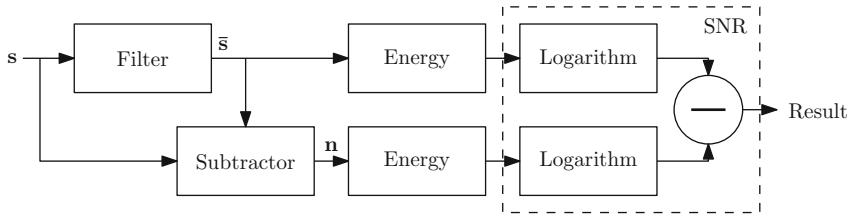


Fig. 22.16 Scheme to compute the SNR

the ECG signal is not available. For sake of generality, we focus on a simple scheme in which the quality of a signal is given by the Signal-to-Noise-Ratio (SNR) between the originally measured signal s and a filtered version of the signal, denoted by \bar{s} , as shown in [6]. The rationale behind this choice is that the filtered (de-noised) signal represents the ideal signal corresponding to s . Furthermore, the removed noise, i.e. the distance of \bar{s} from s , represents a measure of the quality of the measurement s . A natural choice is to represent the SNR, defined as the ratio between the energy of \bar{s} and the energy of the noise $n = \bar{s} - s$, in dB, i.e. by using a logarithmic scale. The reason for such a choice is twofold: first of all this is a consolidated practice in signal processing, moreover we will show that by replacing the computation of a ratio with the computation of the difference of two logarithms, the complexity of a secure protocol for the evaluation of the SNR is considerably reduced. The general structure of the quality evaluation algorithm is reported in Fig. 22.16. We assume that the ECG signal has already been preprocessed to remove known sources of noise such as the interference from the Electrical Grid.

We now describe each block of Fig. 22.16 in details, by describing possible s.p.e.d. implementations of the various blocks.

Filtering: A denoising filter is a linear operator that can be applied in the encrypted domain by using homomorphic encryption as proposed in [10]. An integer low-pass FIR filter, supposed to be intellectual property of \mathcal{S} , is used [6]. The filter, having cut off frequency $f_c = 20$ Hz, is optimized in such a way to minimize the number of coefficients. It has an amplification factor $amp = 64$ and order $k_c = 82$, with only 47 non-zero coefficients, only six of which are non-unitary. The filter coefficients satisfy the following symmetry condition: $c_i = c_{k_c-i}$, hence a filtered sample can be computed as $\bar{s}_i = c_0 s_i + \sum_{j=1}^{k_c/2} (c_j s_{i+j} + c_{j+k_c-i})$. The maximum filter coefficient is 8 and the coefficients assume values in the set $\mathcal{K} = \{0, 1, 2, 4, 5, 6, 7, 8\}$. To filter the signal, each product $\llbracket c_j s_i \rrbracket \forall i = 1, \dots, n \forall j = 0, \dots, k_c/2$ is first computed as

$$\llbracket s_i c_j \rrbracket = \begin{cases} 1 & \text{if } c_j = 0, \\ \llbracket s_i \rrbracket & \text{if } c_j = 1, \\ \llbracket s_i \rrbracket^{c_j} & \text{else.} \end{cases} \quad (22.13)$$

Since there are only six non-null and non-unitary coefficients, only $6k$ products are computed. Then the filtered samples are obtained as

$$\llbracket \bar{s}_i \rrbracket = \llbracket c_0 s_i \rrbracket \prod_{j=1}^{k_c/2} (\llbracket c_j s_{i+j} \rrbracket \llbracket c_j s_{i-j} \rrbracket) \quad \forall i = 1, \dots, k. \quad (22.14)$$

Noise Estimation: The noise signal \mathbf{n} is estimate by subtracting \mathbf{s} and $\bar{\mathbf{s}}$ sample-wise ($n_i = s_i - \bar{s}_i$). Since the integer filter is applied in the encrypted domain, the amplification factor amp must be considered. Thus, to correctly derive the noisy part of the signal, \mathbf{s} must be multiplied by the same factor, obtaining $\llbracket n_i \rrbracket = \llbracket amp * s_i - \bar{s}_i \rrbracket = \llbracket s_i \rrbracket^{amp} * \llbracket \bar{s}_i \rrbracket^{-1} \forall i = 1 \dots k$.

Energy Computation: To compute the SNR the energy of $\bar{\mathbf{s}}$ and \mathbf{n} must be evaluated. This can be done by using an interactive protocol⁹ computing $\llbracket E_x \rrbracket = \llbracket \sum_{i=1}^n x_i^2 \rrbracket = \prod_{i=1}^n \llbracket x_i^2 \rrbracket$, applied to both $\bar{\mathbf{s}}$ and \mathbf{n} , that returns $\llbracket E_{\bar{s}} \rrbracket$ and $\llbracket E_n \rrbracket$.

SNR Computation: The final step of the quality evaluation procedure is the computation of the SNR:

$$SNR = 10 * \log_{10} \frac{E_{\bar{s}}}{E_n} = \frac{10}{\log_2 10} * \log_2 \frac{E_{\bar{s}}}{E_n}. \quad (22.15)$$

Notice that it is easier to compute the difference between two logarithms rather than the logarithm of a ratio. In addition, the factor $10 / \log_2 10$ is only a constant and therefore can be neglected. As a result, the SNR can be computed securely as:

$$SNR = \log_2 \frac{E_{\bar{s}}}{E_n} = \log_2 E_{\bar{s}} - \log_2 E_n. \quad (22.16)$$

While (22.16) is an exact equality when dealing with real numbers, in secure protocols it is only an approximated relationship since an integer logarithm is used. Furthermore, we observe that computing the integer logarithm (base 2) of a binary positive integer number is equivalent to detecting the minimum number of bits necessary to represent the number (in our case the energy is positive by definition). Equation (22.16) can be easily implemented by using a GC.

To switch from HE to GC, the interface described in Sect. 22.4.1.2 can be used: \mathcal{S} obfuscates $E_{\bar{s}}$ and E_n and sends them to \mathcal{C} that, after decryption, uses them as inputs to a GC prepared by \mathcal{S} . The Boolean circuit starts by removing the obfuscations from the energies, then it evaluates the logarithms and finally computes the SNR by subtracting the result of the logarithms. As shown in [6], logarithm computation is performed in two steps: in the first one the input value is processed so that it has zero values before the most significant 1 and a sequence of 1 in the remaining bits, in the second step the bits equal to 1 are counted. This

⁹The computation of the square value needs interaction with the decryption key owner.

can be easily implemented by using a boolean circuit. The protocol is applied to both the energy $E_{\bar{s}}$ and E_n and the output of the first step is denoted with $\omega(E_{\bar{s}})$ and $\omega(E_n)$. Then the *SNR* is simply computed by counting the bits equal to 1 in both $\omega(E_{\bar{s}})$ and $\omega(E_n)$ through a COUNT circuit [6] and then subtracting the two values. Note that the protocol returns $\lfloor \log_2 E \rfloor + 1$, but after the subtraction the two +1 terms are discarded.

A more efficient implementation can be obtained as follows. Given the two energies $E_{\bar{s}}$ and E_n the first part of the protocol is applied to both values obtaining $\omega(E_{\bar{s}})$ and $\omega(E_n)$. The *SNR* is obtained by evaluating $\text{COUNT}(\omega(E_{\bar{s}}) \oplus \omega(E_n))$, in fact the result of the xor is a binary string containing a number of 1's equal to the result of the difference.

The only thing that still needs to be done is the computation of the *SNR* sign, for which it is sufficient to evaluate the relation $E_{\bar{s}} < E_n$ which has the same complexity of a subtraction circuit.

22.6.1.1 Protocol Complexity

We now give an estimate of the complexity of the SNR protocol outlined above, by considering the optimization in the final part. We start by considering the number of bits for an accurate representation of the involved signals and values. In the Physiobank database each sample s_i is represented with $\ell_s + 1 = 11$ bits, hence the maximum value that the samples can assume is $\max_s = 2^{10} - 1 = 1023$. We consider the computation of the SNR of 30 s of signal, sampled at 360 HZ (for a total of $k = 10,800$ samples).¹⁰ Given these parameters, we can compute the bitsize of all the data involved in the computation as shown in Table 22.9.

In the filtering phase a high number of ciphertexts is transmitted. The communication complexity can be reduced by using the composite signal representation

Table 22.9 Number of bits necessary to represent the values obtained by a worst case analysis

Variable name	Maximum value	Magnitude bitlength
Original sample s_i	1023	$\ell_s = 10$
Filter coefficients c_j	8	$\ell_c = 4$
Filtered sample \bar{s}_i	114,576	$\ell_{\bar{s}} = 17$
Noise signal n	180,048	$\ell_n = 18$
Signal energy $E_{\bar{s}}, E_n$	350, 106, 648, 883, 200	$\ell_E = 49$
SNR SNR	49	$\ell_{SNR} = 6$

¹⁰This choice is motivated by the fact that 30 s are sufficient for SNR evaluation and do not introduce a big delay in further computation.

Table 22.10 SNR protocol data transfer

Section	Bits
HE	5,393,856
\mathcal{C} input secret transmission (online OT)	68,480
\mathcal{C} input secret transmission	34,240
Garbled table transmission	149,040
Total	5,645,616

of [10], i.e. packing the ECG samples in n_p ciphertexts. The communication complexity of the SNR protocol for short-term security [6] is summarized in Table 22.10.¹¹

22.6.2 SNR-Based Quality Evaluation

The final assessment of ECG quality is carried out by computing a block-based version of SNR [37]. Specifically, the signal is subdivided into small segments and, for each of them, we compute the SNR as indicated in the previous section. Finally, the mean and variance of the segment-wise SNR are computed and used together with the SNR of the whole signal to make a final decision. The use of the variance is justified by the observation that while the electrode contact noise has a minor impact on the mean SNR, an occasional burst of noise of small length can be better detected by examining the SNR variance.

In the following, we describe the steps necessary to reach a decision based on the following set of features (SNR, segment-SNR mean, segment-SNR variance) extracted during the privacy preserving protocol. The plain implementation of the protocol is summarized in Fig. 22.17. In the remainder of this section, we assume to evaluate κ seconds of an ECG signal $\mathbf{s} = \{s_1, \dots, s_{\kappa f_s}\}$, where f_s is the sampling frequency and each sample is represented with $\ell_s + 1$ bits (ℓ_s bits for the magnitude and 1 for the sign).

Filtering and noise computation: The protocol starts as described in Sect. 22.6.1.

The signal $\mathbf{s} = \{s_1, \dots, s_{\kappa f_s}\}$ is encrypted by \mathcal{C} and transferred to \mathcal{S} . The encrypted signal $[\![\mathbf{s}]\!]$ is filtered by \mathcal{S} to produce the signal $[\![\bar{\mathbf{s}}]\!]$. Finally \mathcal{S} computes the encrypted noise signal $[\![\mathbf{n}]\!]$, i.e. the encryption of the difference between the signals \mathbf{s} and $\bar{\mathbf{s}}$, still using HE.

Energy and SNR Evaluation: The signals $[\![\bar{\mathbf{s}}]\!]$ and $[\![\mathbf{n}]\!]$ are subdivided into segments of w samples, obtaining $m = \lfloor \kappa f_s / w \rfloor$ segments $\mathbf{fs} = \{\mathbf{fs}^1 \dots \mathbf{fs}^m\}$ and $\mathbf{fn} = \{\mathbf{fn}^1 \dots \mathbf{fn}^m\}$, where each segment is composed by the encryption of w samples. For simplicity, we assume that $\kappa f_s = mw$. For each pair of signal

¹¹The communication complexity estimate is performed by considering $T = 1248$ (instead of $T = 1024$, as in the original paper), garbled row reduction and OT precomputation.

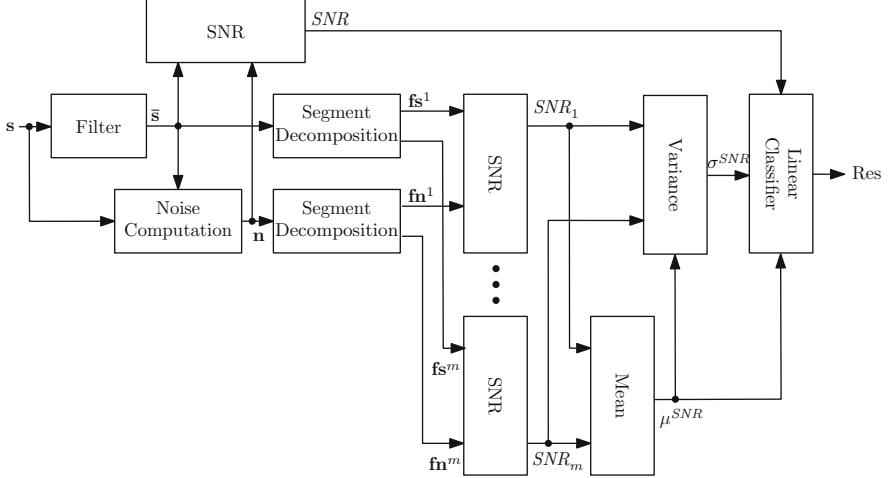


Fig. 22.17 Sequence of steps performed to evaluate the quality of an ECG signal

and noise segments ($\mathbf{fs}^i, \mathbf{fn}^i$) the SNR is evaluated by using the hybrid protocol described in Sect. 22.6.1, obtaining SNR_i^f as garbled secrets. The SNR^f values are not the final result, in fact they have to be used in later computation and hence kept secret. To avoid their disclosure to \mathcal{C} , the GC subprotocol blinds them by adding random values, each $\ell_{SNR_f} + \tau$ bits long, where ℓ_{SNR_f} is the bitlength of the segment SNR.

We underline that the computation of the SNR of the whole signal can be obtained starting from the energy of the segments previously computed. In fact

$$\begin{aligned} SNR &= \log_2 \frac{\sum_{j=1}^{mw} (fs_j)^2}{\sum_{j=1}^{mw} (fn_j)^2} = \log_2 \frac{\sum_{i=1}^m \sum_{j=1}^w (fs_j^i)^2}{\sum_{i=1}^m \sum_{j=1}^w (fn_j^i)^2} \\ &= \log_2 \left(\sum_{i=1}^m E_{fs^i} \right) - \log_2 \left(\sum_{i=1}^m E_{fn^i} \right). \end{aligned}$$

\mathcal{C} adds together all the segment energies E_{fs^i} and all the segment energies E_{fn^i} , obtaining the energies E_x and E_n of the whole signals, obfuscated by a value that is the sum of the obfuscations introduced in each segment, still known by \mathcal{S} . Finally, SNR is obtained by another circuit for SNR evaluation having a larger input bit-length and hence a larger number of gates.

Mean and Variance computation: Once all the obfuscated SNR_i^f and SNR values are obtained, \mathcal{C} encrypts them and transmits the ciphertexts to \mathcal{S} , who can remove the obfuscation by using the homomorphic properties of the cryptosystem and compute their mean and variance by using HE.

The computation of the SNRs mean would require a division by m . If m is public (or known), this is a cheap operation. However, if m is a private value, the

division operation can be expensive and requires an interactive protocol [35, 48]. An additional disadvantage is that this protocol would introduce a rounding error. An alternative is to avoid the division by m . Then, the mean is amplified by a factor of m , the accuracy is preserved and the complexity of the protocol is reduced since interaction with \mathcal{C} is not necessary. The amplified mean is simply computed by \mathcal{S} as $\llbracket \mu^{SNR} \rrbracket = \llbracket \sum_{i=1}^m SNR_i^f \rrbracket = \prod_{i=1}^m \llbracket SNR_i^f \rrbracket$.

The computation of the variance can be performed by using an interactive protocol. Given the amplified mean, the SNR^f values must be amplified by the same factor before computing the variance. Moreover, following the same approach used for the mean, division by m is avoided and thus, the resulting variance is amplified by a factor m^3 . Considering that \mathcal{C} already has the SNR_i^f values obfuscated with a random value r_{SNR_i} , \mathcal{S} can send the obfuscated SNR mean $\llbracket \mu^{SNR} + r_\mu \rrbracket$ (where $r_\mu \in \mathbb{Z}_{\ell_\mu + \tau}$, given the bitlength ℓ_μ of μ^{SNR}) to \mathcal{C} who decrypts it and computes

$$\sum_{i=1}^m \left(m(SNR_i^f + r_{SNR_i}) - (\mu^{SNR} + r_\mu) \right)^2.$$

Finally \mathcal{C} encrypts the result and sends it back to \mathcal{S} who removes the obfuscation value $\sum_i (m r_{SNR_i} - r_\mu)^2 - 2 \sum_i (m SNR_i^f - \mu^{SNR})(m r_{SNR_i} - r_\mu)$ thanks to HE properties, thus, obtaining the encrypted amplified variance $\llbracket \sigma^{SNR} \rrbracket$.

Classification: Each feature previously computed (the overall SNR, the segment-SNR mean and the segment-SNR variance) can be used in a single-feature classifier that compares the feature with a threshold obtained in a training phase. In [37] it is shown that by combining the three features, the accuracy of the classification is higher than by using a simple linear classifier. Specifically, the signal quality is classified by evaluating the following inequality:

$$a + b \sigma^{SNR} + c \mu^{SNR} + d SNR > 0$$

where the coefficients a, b, c, d are obtained by training. If the inequality is verified, the signal is classified as noisy and ECG classification is not carried out.

In the tests, training was carried out in the plain domain separately for each subject by using ECG recorded in a supervised environment so that the weight vector $\mathbf{w} = [a, b, c, d]$ is obtained, with the same procedure described in (22.4). Because $[a, b, c, d]$ are real values, they are quantized and represented with integer numbers. We also observe that the scalar product can be implemented by resorting to HE. In short, signal quality classification is carried out by computing:

$$\llbracket a + b \sigma^{SNR} + c \mu^{SNR} + d SNR \rrbracket = \llbracket a \rrbracket \llbracket \sigma^{SNR} \rrbracket^b \llbracket \mu^{SNR} \rrbracket^c \llbracket SNR \rrbracket^d.$$

The sign of the scalar product is used for the classification, while the magnitude gives an indication of its reliability. High values are more reliable than small

ones. Noticing that both sign and magnitude are useful for the classification and depending on the privacy requirements of \mathcal{S} and \mathcal{C} , the protocol can choose to disclose the scalar product result to \mathcal{C} . Hence, \mathcal{S} sends the encrypted magnitude and sign values to \mathcal{C} who decrypts them and performs the final classification (determines to which class the particular measurement belongs). If \mathcal{S} prefers to keep the magnitude secret, it can obfuscate the scalar product by using a random value and then transmit a comparison GC having the obfuscated value and the random value as inputs.

22.6.2.1 Complexity Analysis

The MIT-BIH Arrhythmia Database¹² has been used in the experiments. The signals were divided into intervals and classified as noisy or clean according to the annotation available in the database, even if the type of noise was not explicitly specified. An interval is considered clean if no samples are affected by noise, otherwise it is considered as noisy. The data set has been extended by adding artificial electrode contact noise stored in the MIT-BIH Noise Stress Test Database¹³ to whole clean segments and partially (only to a randomly chosen section of the segment).

To evaluate the quality of the signal, $\kappa = 30\text{ s}$ of the signal have been analyzed. The signal and the filter have the characteristics outlined in Sect. 22.6.1 and Table 22.9. The 30 s of signal ($t * f_s = 10,800$ samples) are subdivided into $m = 30$ segments, each with length equal to one second ($w = 360$ samples).

The bit-length analysis based on a worst-case assumption [37] gives the results reported in Table 22.11. The maximum value that each variable can assume may be easily determined by a logarithm computation.

The communication complexity of the SNR protocol for short-term security [37] is summarized in Table 22.12.¹⁴ The protocol transmits $\sim 4.3\text{ MB}$ of data for the classification of 30 s of ECG signal. We highlight that more efficient protocols can be developed by moving circuit garbling, circuit transmission and \mathcal{S} input secret transmission to the precomputation phase.

22.6.2.2 Classification Performance

From each signal in the dataset, segments are subdivided into clean (c) or noisy (n) according the notation in the Physiobank databases. In a real implementation an expert decides whether the signals used for training are clean or noisy.

¹²<http://www.physionet.org/physiobank/database/mitdb/>.

¹³<http://www.physionet.org/physiobank/database/nstdb/>.

¹⁴We considered $T = 1248$, garbled row reduction and OT precomputation.

Table 22.11 Maximum value and number of bits necessary for the magnitude representation of the variables involved in the computation by worst case analysis

Variable name	Maximum value	Magnitude bitlength
Original signal s	1023	$\ell_s = 10$
Filter coefficients c	8	$\ell_c = 4$
Filtered signal \bar{s}	114,576	$\ell_{\bar{s}} = 17$
Noise signal n	180,048	$\ell_n = 18$
Segment energy E_{fs}, E_{fn}	11,670,221,629,440	$\ell_e = 44$
Segment SNR SNR_f	44	$\ell_{SNR_f} = 6$
Full energy $E_{\bar{s}}, E_n$	350,106,648,883,200	$\ell_E = 49$
SNR SNR	49	$\ell_{SNR} = 6$
SNR mean μ^{SNR}	1320	$\ell_{\mu} = 11$
SNR variance σ^{SNR}	52,272,000	$\ell_{\sigma} = 26$

Another bit is needed for the sign, except for energies and SNR variance

Table 22.12 Bandwidth (bits) required by the protocol

Section	Online
HE	32,450,496
Circuit	2,551,680
C secrets to GC (online OT)	438,080
S secrets to GC	432,320
Total	35,872,576

Due to the short length of the signals in the database (30 minutes), only 60 segments can be extracted from each signals. Once classified them as noisy or clean according to database notations, signals having less than ten clean segments or less than ten noisy segments have been discarded, because the small number of segments can compromise a good training. Three tests have been performed by using different data sets. The first data set (c/n) is composed by only real clean and noisy sections. The second (c/a) data set has been built by using clean sections and sections where artificial noise was added to entire clean sections (a), considering them as noisy sections. The last data set (c/p) is similar to the second one, with noise added only to a randomly chosen interval of corrupted sections (p).

For each signal, 60 % of the segments of the different types were randomly chosen for the training and the others were used for the testing. A different w vector has been obtained for each individual (signal). Table 22.13 shows that the protocol for remote ECG quality evaluation using the linear classifier guarantees quite good classification results ($\sim 85\%$). The table also shows the performance that could be obtained by classifying the signal by using only one of the three features computed, i.e. the mean of the segment SNR, the variance of the segment SNR or the whole SNR. The table contains the mean of the results of all the signals used for the tests. It is clear from these results that the combination of all derived values with a simple

Table 22.13 Performance of the protocol using the linear classifier or a single feature

Test type	Linear classifier	σ^{SNR}	μ^{SNR}	SNR
c/n	0.8490	0.8158	0.7061	0.7325
c/a	0.8365	0.8005	0.8368	0.8234
c/p	0.7377	0.6729	0.6695	0.6666

linear classifier significantly improves the classification results. On the other hand the use of only one feature does not decrease significantly the protocol complexity.

22.7 Conclusion

The need for privacy protection is steadily increasing in our society due to the wide diffusion of online distributed services offered by untrusted parties having a potential access to private information, like users' preferences or other personal data. This need is even more pressing in settings where the information to be protected is related to the health of the users: with the appearance of more and more online medical repositories, it is simple to imagine that in a few years the approach to remote healthcare will be very different from the current one and it is of the utmost importance that manipulation of sensitive data does not compromise users' privacy.

The availability of signal processing algorithms that work directly on the encrypted data is of great help for application scenarios where biomedical signals must be produced, processed or exchanged in digital format. While, in principle, the evaluation of any functionality in the encrypted domain is always possible, the development of efficient schemes that minimize the computational and communication complexity is not trivial, since it requires a joint design of the signal processing and cryptographic aspects of the system.

In this chapter we have shown how s.p.e.d. technology may help to achieve the twofold goal of allowing the processing of biomedical signals while ensuring the privacy of the signal's owner with reasonable communication and computation complexity. This has been possible by considering the classification of ECG signals by relying on STPC constructions and on a proper design of the classification algorithms so as to ease their implementation in a s.p.e.d. framework. More specifically, we have described two ECG classification protocols; the former is based on a Linear Branching Program (LBP) classifier, where Quadratic Discriminant Functions (QDF) or Linear Discriminant Functions (LDF) are used in the nodes, the latter implements a Neural Network (NN) classifier. The optimization of the signal processing part substantially improved the performance of both the solutions.

We also presented a protocol to evaluate the quality of an ECG signal specifically geared towards remote health monitoring applications. The solution analyzes the amount of noise in a biomedical signal, based on analysis of the Signal-to-Noise ratio in small segments of the encrypted signal, rather than the whole measurement.

For both the classification and quality evaluation protocols, a reader may wonder whether a running time in the order of a few seconds is affordable in real life applications. The ultimate answer to this question depends on the application at hand, however we can observe that a running time of less than one second would be enough for applications wherein heart beats are classified at the same pace at which they are produced. While the protocols we have described are not that fast, their performance are not far away from the above so-to-say real time requirements thus encouraging further research.

Acknowledgements We would like to thank the co-authors of the original papers, i.e., Pierluigi Failla, Jorge Guajardo, Vladimir Kolesnikov, Annika Paus, Ahmad-Reza Sadeghi, Thomas Schneider (in alphabetical order), for the important contribution provided to the research in this interesting field.

References

1. Acharya, U.R., Suri, J., Spaan, J.A.E., Krishnan, S.M.: Advances in Cardiac Signal Processing. Springer, Heidelberg (2007)
2. Arulampalam, G., Bouzerdoum, A.: Application of shunting inhibitory artificial neural networks to medical diagnosis. In: Intelligent Information Systems Conference, The Seventh Australian and New Zealand 2001, pp. 89–94 (2001)
3. Asharov, G., Lindell, Y., Schneider, T., Zohner, M.: More efficient oblivious transfer and extensions for faster secure computation. In: Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, Berlin, pp. 535–548. <http://www.sigsac.org/ccs/CCS2013/> (2013)
4. Barni, M., Failla, P., Kolesnikov, V., Lazzeretti, R., Sadeghi, A.R., Schneider, T.: Secure evaluation of private linear branching programs with medical applications. In: 14th European Symposium on Research in Computer Security (ESORICS'09), Saint Malo, LNCS. Springer, Heidelberg (2009). <http://eprint.iacr.org/2009/195> (2009)
5. Barni, M., Failla, P., Lazzeretti, R., Paus, A., Sadeghi, A., Schneider, T., Kolesnikov, V.: Efficient privacy-preserving classification of ECG signals. In: First IEEE International Workshop on Information Forensics and Security, 2009. WIFS 2009, pp. 91–95. IEEE, London (2009)
6. Barni, M., Guajardo, J., Lazzeretti, R.: Privacy preserving evaluation of signal quality with application to ECG analysis. In: 2010 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6. IEEE, Seattle (2010)
7. Barni, M., Failla, P., Lazzeretti, R., Sadeghi, A., Schneider, T.: Privacy-preserving ECG classification with branching programs and neural networks. In: IEEE Transactions on Information Forensics and Security (TIFS) (2011)
8. Beaver, D.: Precomputing oblivious transfer. In: Advances in Cryptology – CRYPTO'95, Santa Barbara. LNCS, vol. 963, pp. 97–109. Springer, Heidelberg (1995)
9. Bellare, M., Hoang, V., Keelveedhi, S., Rogaway, P.: Efficient garbling from a fixed-key blockcipher. In: 2013 IEEE Symposium on Security and Privacy (SP), San Francisco, pp. 478–492 (2013)
10. Bianchi, T., Piva, A., Barni, M.: Composite signal representation for fast and storage-efficient processing of encrypted signals. IEEE Trans. Inf. Forensic Secur. **5**(1), 180–187 (2010)
11. Blanton, M., Gasti, P.: Secure and efficient protocols for iris and fingerprint identification. In: Computer Security–ESORICS 2011, Leuven, pp. 190–209 (2011)

12. Blommestein, H.: Specification and estimation of spatial econometric models: A discussion of alternative strategies for spatial economic modelling. *Reg. Sci. Urban Econ.* **13**(2), 251–270 (1983)
13. Campisi, P.: Security and Privacy in Biometrics, chap. R. Lazzeretti and P. Failla and M. Barni. *Privacy-Aware Processing of Biometric Templates by Means of Secure Two-Party Computation*. Springer, Heidelberg (2013)
14. Coron, J., Mandal, A., Naccache, D., Tibouchi, M.: Fully homomorphic encryption over the integers with shorter public keys. In: *Advances in Cryptology–CRYPTO 2011*, Santa Barbara, pp. 487–504 (2011)
15. Damgård, I., Geisler, M., Krøigaard, M.: Efficient and secure comparison for on-line auctions. In: *Information Security and Privacy*, pp. 416–430. Springer, Heidelberg (2007)
16. Demmler, D., Schneider, T., Zohner, M.: ABY – a framework for efficient mixed-protocol secure two-party computation. In: *21st Annual Network and Distributed System Security Symposium (NDSS’15)*. The Internet Society, San Diego (2015). doi:10.14722/ndss.2015.23113. <http://thomaschneider.de/papers/DSZ15.pdf>. Code: <http://encrypto.de/code/ABY> (2015)
17. ElGamal, T.: A public key cryptosystem and a signature scheme based on discrete logarithms. *IEEE Trans. Inf. Theory* **IT-31**(4), 469–472 (1985)
18. Erkin, Z., Piva, A., Katzenbeisser, S., Lagendijk, R., Shokrollahi, J., Neven, G., Barni, M.: Protection and retrieval of encrypted multimedia content: when cryptography meets signal processing. *EURASIP J. Inf. Secur.* **2007**, 17 (2007)
19. Even, S., Goldreich, O., Lempel, A.: A randomized protocol for signing contracts. *Commun. ACM* **28**(6), 647 (1985)
20. Fontaine, C., Galand, F.: A survey of homomorphic encryption for nonspecialists. *EURASIP J. Inf. Secur.* **2007**(1), 1–15 (2007). doi:<http://dx.doi.org/10.1155/2007/13801>
21. Ge, D.F., Srinivasan, N., Krishnan, S.M.: Cardiac arrhythmia classification using autoregressive modeling. *Biomed. Eng. (online)* **1**(1), 5 (2002)
22. Gentry, C.: Fully homomorphic encryption using ideal lattices. In: *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, Bethesda, pp. 169–178 (2009)
23. Giry, D., Quisquater, J.J.: Cryptographic key length recommendation (2009). <http://keylength.com>
24. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: Physiobank, Physiotoolkit, and Physionet: components of a new research resource for complex physiologic signals. *Circulation* **101**(23), e215–e220 (2000)
25. Goldreich, O.: Secure multi-party computation. Manuscript. Preliminary version (1998). [Http://citeserx.ist.psu.edu/viewdoc/download?doi=10.1.1.11.2201&rep=rep1&type=pdf](http://citeserx.ist.psu.edu/viewdoc/download?doi=10.1.1.11.2201&rep=rep1&type=pdf)
26. Huang, Y., Evans, D., Katz, J., Malka, L.: Faster secure two-party computation using garbled circuits. In: *USENIX Security Symposium*, San Francisco, vol. 201 (2011)
27. Ishai, Y., Kilian, J., Nissim, K., Petrank, E.: Extending oblivious transfers efficiently. In: *Advances in Cryptology – CRYPTO’03*, Santa Barbara. LNCS, vol. 2729. Springer, Heidelberg (2003)
28. Kattan, M., Beck, R.: Artificial neural networks for medical classification decisions. *Arch. Pathol. Lab. Med.* **119**(8), 672–677 (1995)
29. Kilian, J.: Founding crytpography on oblivious transfer. In: *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, Chicago, pp. 20–31 (1988)
30. Kolesnikov, V., Sadeghi, A., Schneider, T.: How to combine homomorphic encryption and garbled circuits. In: *Signal Processing in the Encrypted Domain–First SPEED Workshop–Lousanne*, p. 100 (2009)
31. Kolesnikov, V., Schneider, T.: Improved garbled circuit: Free XOR gates and applications. In: *International Colloquium on Automata, Languages and Programming (ICALP’08)*, Reykjavik. LNCS, vol. 5126, pp. 486–498. Springer, Heidelberg (2008)
32. Kolesnikov, V., Sadeghi, A., Schneider, T.: A systematic approach to practically efficient general two-party secure function evaluation protocols and their modular design. *J. Comput. Secur.* **21**(2), 283–315 (2013)

33. Lagendijk, R., Erkin, Z., Barni, M.: Encrypted signal processing for privacy protection: Conveying the utility of homomorphic encryption and multiparty computation. *IEEE Signal Process. Mag.* **30**(1), 82–105 (2013). doi:10.1109/MSP.2012.2219653
34. Lazzeretti, R.: Privacy preserving processing of biomedical signals with application to remote healthcare systems. Ph.D. thesis, PhD school of the University of Siena, Information Engineering and Mathematical Science Department (2012). <Http://theses.eurasip.org/theses/472/privacy-preserving-processing-of-biomedical/download/>
35. Lazzeretti, R., Barni, M.: Division between encrypted integers by means of garbled circuits. In: 2011 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6. IEEE, Iguaçu Falls (2011)
36. Lazzeretti, R., Barni, M.: Private computing with garbled circuits [applications corner]. *IEEE Signal Process. Mag.* **30**(2), 123–127 (2013)
37. Lazzeretti, R., Guajardo, J., Barni, M.: Privacy preserving ECG quality evaluation. In: 14th ACM Workshop on Multimedia and Security (MM&Sec 2012), Coventry (2012)
38. Lindell, Y., Pinkas, B.: A proof of Yao’s protocol for secure two-party computation. *J. Cryptol.* **22**(2), 161–188 (2009). Cryptology ePrint Archive: Report 2004/175
39. Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Advances in Cryptology—EUROCRYPT ‘99, Prague, pp. 223–238 (1999)
40. Pignata, T., Lazzeretti, R., Barni, M.: General function evaluation in a STPC setting via piecewise linear approximation. In: 2012 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, Tenerife (2012)
41. Pinkas, B., Schneider, T., Smart, N., Williams, S.: Secure two-party computation is practical. In: Advances in Cryptology—ASIACRYPT 2009, Tokyo, pp. 250–267 (2009)
42. Pisa, P., Abdalla, M., Duarte, O.: Somewhat homomorphic encryption scheme for arithmetic operations on large integers. In: Global Information Infrastructure and Networking Symposium (GIIS), 2012, pp. 1–8. IEEE, Choroni (2012)
43. Prabhakaran, M., Sahai, A.: Secure Multi-Party Computation. IOS Press, Amsterdam (2013)
44. Rivest, R., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. *Commun. ACM* **21**(2), 120–126 (1978)
45. Rivest, R.L., Adleman, L., Dertouzos, M.L.: On data banks and privacy homomorphisms. In: Demillo, R.D. et al. (eds.) Foundations of Secure Computation, pp. 169–179. Academic, New York (1978)
46. Sadeghi, A., Schneider, T., Wehrenberg, I.: Efficient privacy-preserving face recognition. In: International Conference on Information Security and Cryptology – ICISC 2009, Seoul (2009)
47. Schoenmakers, B., Tuyls, P.: Efficient binary conversion for Paillier encrypted values. Advances in Cryptology-EUROCRYPT 2006, Saint Petersburg, pp. 522–537 (2006)
48. Veugen, T.: Encrypted integer division. In: 2010 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6. IEEE, Seattle (2010)
49. Yao, A.C.: Protocols for secure computations. In: IEEE Symposium on Foundations of Computer Science (FOCS’82), Chicago, pp. 160–164 (1982)
50. Yao, A.C.: How to generate and exchange secrets. In: IEEE Symposium on Foundations of Computer Science (FOCS’86), Toronto, pp. 162–167 (1986)

Chapter 23

Strengthening Privacy in Healthcare Social Networks

Maria Bertsimas, Iraklis Varlamis, and Panagiotis Rizomiliotis

Abstract The purpose of this chapter is to present a comprehensive approach to the problem of privacy protection in healthcare social networks, to summarize threats and suggest emerging technological solutions to protect users. For this purpose we start with a definition of the term “privacy” and how it evolved through time. We continue within the context of social networks and highlight the main privacy issues and threats for network members. In addition, we analyze the Common Criteria for IT security evaluation that apply to privacy, under the prism of Healthcare Social Networks (HSNs) and present tools and methods that may enhance privacy in such networks. Finally, we provide examples of popular HSNs, categorized according to the purpose they serve and discuss the privacy challenges for them.

23.1 Introduction

Privacy as a concept has been studied in depth by researchers from different disciplines such as philosophy, social sciences and law, but yet has not a commonly accepted definition. The American judges Warren and Brandeis [31] in their article entitled “The Right to Privacy” provided one of the oldest and most popular definitions of privacy as a person’s “right to be left alone”. Alan Westin in his “Privacy and Freedom” book [32] gave a more comprehensive definition of privacy: “Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others.”

Originally the term *privacy* was coined to the protection of the home and in general the physical space surrounding a person. When mail first appeared as a means of communication, the first mail privacy violation phenomena were recorded

M. Bertsimas • I. Varlamis (✉)

Department of Informatics and Telematics, Harokopio University of Athens, Athens, Greece
e-mail: itp13307@hua.gr; varlamis@hua.gr

P. Rizomiliotis

Department of Information and Communication Systems Engineering, University
of the Aegean, Lesbos, Greece
e-mail: prizomil@aegean.gr

as early as 1624. The advent of mass media in the dawn of 1900, in conjunction with inventions such as photography and telephony resulted in the first incidents of breaches in private life and phone conversations. In the second half of the twentieth century, the appearance of personal computers and computer networks gave a rise to concerns about privacy issues and turned focus to data collection and processing [17]. It becomes obvious that depending on the data communication, collection, storage and processing methods that existed in every era, respective mechanisms have been developed for violating the privacy of individuals. Thus, the need for privacy has created the need for legislative mechanisms for protection.

Privacy has been included in the Universal Declaration of Human Rights issued by the United Nations in 1948. According to Article 12 of the Declaration:

No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.

Finally, according to Articles 7 and 8 of the Charter of Fundamental EU rights:

Everyone has the right to respect for his or her private and family life, home and communications... and ... the right to the protection of personal data concerning him or her. ... Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data, which has been collected concerning him or her, and the right to have it rectified. ... Compliance with these rules shall be subject to control by an independent authority.

Although privacy cannot be restricted only to personal data, since it refers to the general “right to solitude and withdrawal” [21], the prevention from “unauthorized use or disclosure of data” [27] became the main objective of privacy protection devotees.

A healthcare social network allows its members such as doctors, patients and caregivers to communicate and collaborate in order to virtually manage the illnesses and improve the quality of patients’ life. In advice seeking social networks, patients and caregivers seek for advice on patient care or share experiences and problems. Physicians on the other side, answer questions (e.g., in MedHelp or WellSphere), provide emotional support and contribute their medical experience (e.g., in DailyStrength or OrganizedWisdom). In patient-centric social networks (e.g., PatientsLikeMe, SugarStats) patients share their personal medical records to perform self-tracking. Finally, healthcare professionals join on-line social networks (e.g., Sermo) seeking for best-practices, job openings and career tips, research and product information, as well as the opportunity to securely communicate with their peers. Figure 23.1 illustrates the various participants of a HSN and the information they exchange. Professional networks (in yellow background) attract professionals, researchers, companies and institutions, and disseminate medical information. Patient-centric networks (in blue background) attract patients, volunteers and professionals that wish to support patients. Physicians and nurses usually participate in both types of networks.

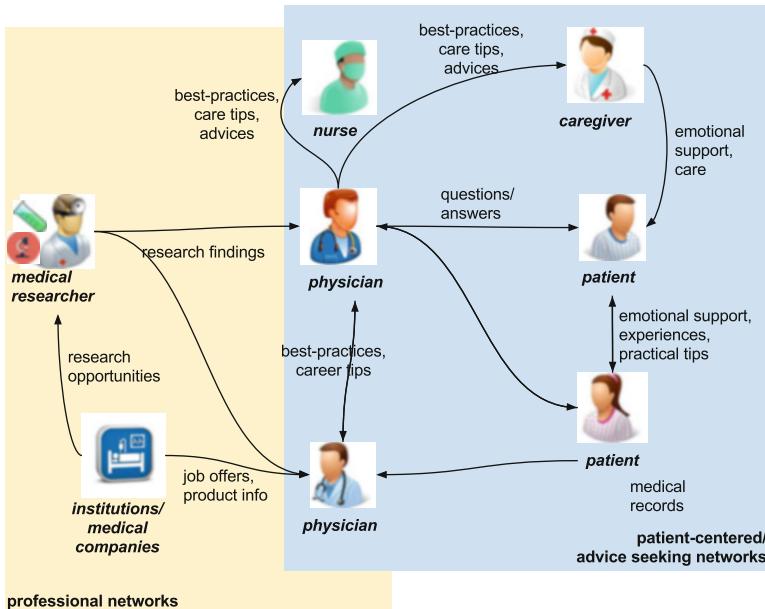


Fig. 23.1 The participants of healthcare social networks

There is not always a single way to classify a healthcare social network as advice-seeking, patient-centric or professional and there are many HSNs that provide services that cover the needs of patients and physicians. An interesting survey of the different types of HSNs is provided by Melanie Swan [28].

23.2 Social Networks

The term social network was invented in 1954 by LSE's Professor JA Barnes [2] who examined social relationships in a Norwegian fishing village, concluding that the entire social life of the village could be represented as "a set of points, some of which are connected by lines", to form a "comprehensive network" of relations.

23.2.1 On-line Social Networks

Recently, the advent of the so-called social networking sites attracted a large number of users and brought the concept of "on-line social networks" in light. The definition given by Ellison et al. [8], clearly defines the assets of a user that joins an on-line social network:

Social network sites are web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system.

As a result, when users join an on-line social network, such as Facebook, Twitter, LinkedIn, Google+, etc. to mention some of them, they are invited to fill their profile page by providing a range of personal information, then encouraged to create digital relationships with other members of the network who either already know or want to know and finally, are asked to maintain these relationships. As a result, they agree to share profile and connections with other (selected) users.

In a broader definition, given by the European Agency for Information Systems Security (ENISA) [16], on-line social networks are defined as federated identity management spaces, where users: (a) store and manage personal data, (b) control access to them based on credentials, (c) are able to find out who has accessed their personal data.

For example, a user in a social network can allow access to her personal information to her friends, to friends of her friends, or to a specific user group and can never be sure about the real identity of an on-line user. In the case of Healthcare Social Networks, a doctor must be able to access private patient information, only if he can prove that he is really a doctor and he has been granted access by the user. Doctor's credentials (e.g., digital certificates) must be granted during registration and after off-line verification (e.g., a phone call, or an application at the medical association), although this is not always the case.¹ The patient must be able to grant access to her personal information at an individual user level, e.g., to her doctor only and must also be notified when someone accesses her private data.

23.2.2 *Healthcare Social Networks*

In a general context, we could say that social networks can take different forms, depending on the purpose they serve. Although general purpose social networks are the most popular among them, we can also find professional (business) networks, network that relate to health (healthcare), and others that are designed with ethnic, religious or political criteria [37]. Healthcare professionals use traditional social media networks to connect with others, but also 1 out of 3 joins social networks to seek for a job (36 % of them are allied professionals, 33 % are nurses, 29 % are pharmacists and 23 % physicians).²

¹In autumn 2007, a weak security model has been reported for Sermo network, which allowed non-physician applicants to register as physicians using false identities from the Internet.

²Infographic provided by MEdTech Media at: http://www.medtechmedia.com/files/medtech_images/Infographic_SOCIAL_MEDIA_SURVEY_AMN_HEALTHCARE.jpg

According to a more recent study carried in Greece [1], the percentage of doctors and nurses that join social networks for professional use is similar (37 %). However, the percentage for non-medical personnel (e.g., hospital administrative staff) is higher (54 % of them joins social networks for professional use). In this same study, authors detected three groups of users: (a) users aged 22–44 with a few years in service (<10 years) who are enthusiasts of using social media both in their social and professional activities, (b) users of the same age with more years in service (11–20 years) who are still sceptical on using and trusting social networks and (c) seniors (>45 years old) with an extended professional experience (>20 years), but low familiarity with new technologies who are more reluctant in adopting them.

Gunther Eysenbach used the terms “infodemiology” and “infoveillance”, in order to describe a new emerging approach for public health [9], based on large-scale monitoring and data mining of information published in (health and general purpose) social networks. Healthcare social networks allow their members such as doctors, patients and caregivers to communicate and collaborate in order to manage the illnesses and improve the quality of patients’ life. The on-line environment removes distance and time barriers, enables patients to submit on-line requests for advice and share problems and solutions with other patients and facilitates doctors to cooperate with each other and supervise their patients. In analogy to social networks, users in on-line healthcare networks must be ready to share their personal data, which in this case are sensitive health data, with other network members.

In order for the healthcare network to thrive, members need to trust each other and be confident for the secure, reliable and lawful operation of the network [4]. On-line healthcare networks have some unique characteristics, which make the aforementioned targets hard to accomplish. They cross national borders and operate in a continuous basis; they are responsible for securing members’ medical data and are entrusted to preserve members’ anonymity. At the same time, they must guarantee the reliability of both participating members and submitted content.

Usually members’ anonymity is achieved via pseudonyms (virtual identities). This allows the identification of patients within the community and the association to their data. Similarly the use of pseudonyms or real identities allows physicians to clearly identify themselves to lend credibility and accountability to their on-line communications. However, in the former case it remains to the community owners to certify the real identity and expertise of doctors that use pseudonyms. Under these circumstances, the smooth operation of such social networks is a heavy duty for moderators and administrators.

The concept of privacy is strongly connected to the (social and technological) trust of users to the network and the application provider. Users are willing to upload their data to the on-line healthcare application, only when they are confident about the correct use of their data, from the authorized people and for the appropriate purpose. According to the study of Grajales et al. [15] a 76 % of the social media users that participated believe that their personal health data could be used without their knowledge either in order to deny them healthcare benefits (72 %) or to deny them job opportunities (66 %). According to another study by Damschroder et al. [6] patients are positive in sharing their medical data with experts only when they

trust that they will be kept private and confidential in a way that patients can see and understand; when clear and consistent consequences exist for privacy violations and when the computerized systems are proven to be secure.

23.3 Privacy

In the following, we provide an introduction to the concept of privacy, starting with the evolution of its meaning over the ages and focusing on how this meaning was adapted to on-line healthcare networks. Then we distinguish between personal and sensitive data, discuss the legal framework that rules their use, and present the principles that must govern privacy protection and personal data. Finally, we provide a list of potential privacy threats that apply in HSNs.

23.3.1 Background

Nowadays, the use of Internet has enabled the instant diffusion of information and has transformed on-line social networks to repositories, where personal data is collected, enriched, modified, shared and reused continuously [23]. Consequently, the concept of privacy has been transformed from one's "right to be left alone" [21] to the ability of social network users "to control and protect personal information" [27].

According to Rosenberg [24] privacy can be: (a) *territorial*, when it refers to the physical space of an individual; (b) *personal*, when it protects the individual from unwanted interventions; (c) *informational*, when it defines how personal data are collected, stored and processed and who can gain or grant access to. A more recent approach by Finn et al. [10] distinguishes seven types of privacy:

1. The privacy of the person.
2. The privacy of behaviour and action.
3. The privacy of data and image.
4. The privacy of communication.
5. The privacy of thought and feelings.
6. The privacy of space and location.
7. The privacy of association and group privacy.

Although, Finn's work focuses on privacy and user identification, it provides a useful list of user-identification technologies and their impact to the different aspects of privacy. It is interesting that some of the identification technologies mentioned in the study (listed in Table 23.1), are strongly related to medical issues (e.g., body scanners, DNA sequencing and next generation biometrics) and definitely collect information that must be privacy protected.

Table 23.1 Identification techniques that may affect medical privacy (based on [10])

Technology	Whole body imaging scanners	Second generation DNA sequencing	Second generation biometrics
Type of privacy			
Privacy of the person	X	X	X
Privacy of behaviour and action	X	X	X
Privacy of communication			X
Privacy of data and image	X	X	X
Privacy of thought and feelings			X
Privacy of location and space		X	X
Privacy of association		X	X

This example is an evidence that even the participation in on-line healthcare networks must begin with private information sharing.

Even when more simple identification mechanisms are used in on-line healthcare networks, users share their medical records in order to receive medical advices from experts, to monitor the progress of their health, etc. In this context, users reveal their data to specific users (e.g., their doctors) and hide them from others (for example from insurance companies or their employers). Revealing medical data could also reveal medical or psychological conditions, treatments or other details about personal life. Privacy in healthcare networks may comprise control over personal information (informational privacy), physical restriction to data accessibility (physical security) and the respect of the doctor to patients' beliefs, thoughts, values and feelings (psychological security—confidentiality)[26].

23.3.2 Personal and Sensitive Data

“Personal data” and “sensitive data” are the two main assets of HSN users, which must be protected or controlled within a privacy context.

According to the European Data Protection Supervisor (EDPS) glossary and Article 2 (a) of Regulation (EC) No 45/2001 the term “*personal data*” may refer to:

... any information relating to an identified or identifiable natural person (referred to as “data subject”), in particular by reference to an identification number or to one or more factors specific to his or her physical, physiological, mental, economic, cultural or social identity.

The name and the social security number are two examples of personal data, which relate directly to a person. But the definition also extends further and also encompasses for instance e-mail addresses and the office phone number of an employee. Other examples of personal data can be physical characteristics, education, labour, economic status, interests, activities, habits or any other information found in the medical records of a patient or in the evaluation report of an employee.

The term “sensitive data” is coined with “information that reveals racial or ethnic origin, political opinions, religious or philosophical beliefs, membership of a trade union, health status, social welfare, erotic life preferences, prosecutions and convictions, etc.” (Article 10 of Regulation 45/2001; Article 8 of Directive 95/46/EC). Usually, sensitive personal data are legally protected by more stringent regulations than simple personal data. The processing of such information is in principle prohibited, except in specific circumstances. It is possible to process sensitive data for instance if the processing is necessary for the purpose of medical diagnosis, or with specific safeguards in the field of employment law, or with explicit consent of the data subject.

Privacy is a broader concept than personal data protection, since the former includes also “the right to be left alone, out of public view, and in control of information about oneself”. However, personal data protection arises as a necessity to ensure privacy and therefore constitutes a part of the privacy concept. As a result, personal data protection is a prerequisite for guaranteeing privacy against potential technological threats. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) recognized the possible threat to privacy from the electronic data exchange in healthcare. HIPAA mandated the U.S. Department of Health and Human Services (HHS) to develop regulations that protect privacy and security of electronically-transmitted health information.

In 2000, HHS created a set of rules known as the Privacy Rule, which sets a framework for handling medical records and other sensitive personal health information. Privacy Rule defines safeguards, uses and disclosures and patient authorizations and provides patients with certain rights over their health information. In order to be properly applied, the Security Rule requires appropriate administrative, physical and technical safeguards to ensure the confidentiality, integrity, and security of sensitive electronic health information. However, both HIPAA Privacy and Security Rules are designed to establish minimum standards for the purpose of setting the “legal floor”.

Although, HIPAA and Privacy Rule cover the cases of health data stored in electronic health records at medical facilities, this is not guaranteed for HSNs, which are not HIPAA-covered entities. According to Grajales et al. [15]:

... the same information when shared on a social networking site does not have the same protection, and potentially may lead to discrimination by a future employer or health insurance agency, although this has not been reported in the literature to date.

The Health Information Technology for Economic and Clinical Health (HITECH) Act was created in 2009, to stimulate the adoption of electronic health records (EHR) and supporting technology in the United States. HITECH promotes the meaningful use of EHRs and penalizes any improper use. HITECH and HIPAA are separate laws that reinforce each other in certain ways. For example, HITECH requires that any physician and hospital must perform a HIPAA security risk assessment before gaining access to patient HER. In addition, HITECH establishes data breach notification rules and HIPAA holds healthcare providers’ business associates equally accountable for data breaches to the providers themselves.

The Personal Information Protection and Electronic Documents Act (PIPEDA) was the Canadian answer to HIPAA that became a law in 2000. PIPEDA governs how private sector organizations collect, use and disclose personal information in the course of commercial business. The PIPEDA compliance laws state that personal information must be protected by security safeguards appropriate to the sensitivity of the information, including technological measures, such as the use of passwords and encryption. However, according a research from [3], the tested companies lack at best, and entirely neglect at worst the data disclosure compliance issue in relation to the data they collect.

23.3.3 *Privacy Principles*

The practices used by social networking sites like Facebook, Twitter or Instagram and the various stories concerning privacy violations in such networks, make health social network members sceptical about sharing personal data, or even posting questions. However, they are prone to share data on specialized sites, with a transparent security and privacy policy and healthcare professionals that they can trust.

The protection from unauthorized access is a step towards personal data protection. However, more principles must be guaranteed in order to achieve the protection of personal data. In this context, the Organisation for Economic Co-operation and Development (OECD) defined in 1980 the principles that must govern privacy protection and personal data exchange across countries. These principles are:

1. **Collection Limitation Principle:** The collection of personal data should be made using fair and lawful means, and—where possible—with the consent of the data subject. For example, in a healthcare network, patients have the right to exercise control over what medical data to share and with whom.
2. **Data Quality Principle:** Personal data should be relevant to the purpose for which they are to be used and depending on the extent necessary for this purpose should be complete, accurate and updated. In HSNs, members must be able to decide upon the amount and detail of medical data they share (e.g., full history, or selected records) depending on the purpose they will be used.
3. **Purpose Specification Principle:** The purposes for which personal data are collected should be specified before data collection and any subsequent use must be limited to the fulfilment of those objectives or some fully compatible objectives. The HSNs' members must be informed on the purpose for which they share data.
4. **Use Limitation Principle:** Personal data will not be communicated and made available to third parties or used for purposes other than specified, unless the data subject agrees or the law authorizes such changes. Members must be asked for any future use that is beyond the original purpose of data sharing (e.g., a future population analysis for a different purpose).

5. **Security Safeguards Principle:** Personal data should be protected using appropriate mechanisms against risks such as unauthorized access, destruction, use, modification or disclosure to third party entities. The principle applies to HSNs in a straightforward way, in order to guarantee the safe storage of medical data.
6. **Openness Principle:** There should be a general openness policy regarding the practices that relate to the collection and processing of personal data and the identity of the body performing the collection and processing. The organization behind the HSN must clearly state its identity and purposes of the network. When a member wants to share his/her data to third parties, the network must be able to provide access to data.
7. **Individual Participation Principle:** Each person should have the right to:
 - a. Receive a confirmation from the data controller administrator, that person related data are in the data controller's possession.
 - b. Receive information on the data that is of interest to him/her in a reasonable time limit, in an understandable manner and low price (if any).
 - c. Be informed for the reasons that he/she cannot have a and b above and argue, question and further claim these rights.
 - d. Correct or delete personal data.

As a result, HSN's members become the real owners of their own data.

8. **Accountability Principle:** Any personal data controller should be accountable regarding the implementation of those measures that promote the aforementioned principles, which should govern the protection of personal data. As a result, the organization behind an HSN (e.g., a hospital) is responsible for any loss, leak or damage of sensitive information.

It is worth noting that the definition of the above principles was a first step towards establishing rules for the management of personal data and has influenced the legislation that currently governs the protection of personal data. The above principles highlight the strong connection between the concepts of privacy and protection of personal data. Table 23.2 summarizes the Privacy Principles described by OECD and their application in HSNs.

23.3.4 Privacy Threats

According to ENISA [16] the risks for social network users can be categorized as follows: (a) privacy related threats, (b) traditional risks of internet that embrace the environment of social networks (SNS variants of traditional network and information security threats), (c) risks associated with the user ID (Identity related threats), and (d) social risks (social threats).

The most important threats for privacy, under the prism of healthcare social networks, are summarized in Table 23.3 and are discussed in detail in the following paragraphs.

Table 23.2 Privacy principles and their application in HSNs

Privacy principle	Application in HSNs
Collection limitation	The HSNs members can control what medical data the network collects from them for a purpose
Data quality	The members can limit the amount or detail of data collected or retained for them
Purpose Specification	The members must be aware of the purpose for sharing their medical data with the community
Use limitation	The members must be asked for any future use of their data
Security safeguards	Medical data must be safe kept
Openness	A member's data can be made available to other networks or third parties upon member's request
Individual participation	A member can freely act upon his/her data, share with others, provide or deny access
Accountability	The HSNs owners are primarily responsible for any loss, leak or damage of data

Table 23.3 A list of privacy threats and associated risks in HSNs

Threat	Risk
Digital dossier aggregation	Information can become embarrassing or even damaging
Difficulty of complete account deletion	The user loses control over his/her identity. Damaging information cannot be removed, increasing the "digital dossier" effect
Secondary data collection	Disclosed information used for personalization or other purposes in a network, can also be used for targeting or discrimination or resaled to third parties
De-anonymization attacks	The true user identity may be revealed
Inference attacks	Parts of a user identity may be revealed
Identity theft	Can possibly harm the reputation or credibility of the real owner's profile
Phising	The attacker may collect reliable information on his victims and expose it or sale it
Communication tracking	User profile is created by monitoring user's communications. Then the information can be exposed or exploited ("digital dossier" effect)
Information leakage	Medical information may be exposed accidentally or in purpose to unauthorized third-parties

23.3.4.1 Digital Dossier Aggregation

On-line social networks are an ideal source for user profile information. Modern technology enables the automated aggregation of user profiles data and the storage of all traces in order to create a digital user dossier. Such data can be used for purposes other than those for which the user intended, or may take different meaning

outside the initial healthcare network context. More specifically, users publish to healthcare social networks information, which is intended for a specific audience and may prove embarrassing for them when it goes outside this audience.

23.3.4.2 Difficulty of Complete Account Deletion

When a user leaves the social network, he/she usually wants to delete the profile data and any other digital traces or to be able to take them off-line. However, network providers cannot always guarantee that all traces are completely erased [16]. This persistence of user data is yet another privacy risk. For example, even when data are deleted from the user profile, other data such as messages or comments exchanged with other users are not removed from the pages of those users. In general, there is ambiguity about whether indeed the user information in social networks is permanently deleted or copies of user data are kept in storage. The website justdelete.me keeps an up-to-date list of the most popular sign-on websites and describes the process for deleting user profile data. However, there is not yet a HSN in the list.

When data are stored in third-party cloud storage services, then the ambiguity about complete account deletion is stronger. According to U.S. Department of Health and Human Services,³

Covered entities may disclose protected health information to an entity in its role as a business associate only to help the covered entity carry out its health care functions not for the business associate's independent use or purposes, except as needed for the proper management and administration of the business associate.

This means, that health information is stored properly on a third-party cloud-based solution as long as the company has signed a business associates⁴ agreement (BAA) with HHS. However, in the majority of HSNs the underlying agreements are not revealed.

23.3.4.3 Secondary Data Collection

In secondary data collection, the attacker collects information about users of a social network using secondary sources rather than the network itself. The attacker can be any individual who wants to collect more (sensitive) information about another member of the network. Specifically, the attacker uses sources (e.g., search engines) outside the social network for the collection of a user's data and then links these data with the profile of that user. In this way, the attacker is able to gather the maximum

³www.hhs.gov.

⁴Health information privacy: business associates. Available at <http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/businessassociates.html>. Accessed June 5, 2015.

information about the user by any available Internet source. Such techniques can be applied in cases where users keep their profiles private. Then through secondary data collection, the person concerned can collect all possible information from other internet sources [5]. The attacker can be any individual who wants to breach other persons' privacy with the objective to collect information, even identification information (e.g., the social security number or the Foedselsnummer in Norway), which often acts as a key to access personal information from a wide range of different sources.

23.3.4.4 De-Anonymization Attacks

In many social networking sites, especially in those that focus in healthcare, users want to protect their privacy and their anonymity by using virtual identities. At the same time, they upload personal data, which if are properly collected, analysed and combined with information from other sites [20] may reveal the true user identity.

De-anonymization attacks aim to reveal user identities and expose them to the members of the network or to the public. In [25] authors present an approach to disambiguate extracted identity information relating to different individuals through the use of their circle-of-friends. Authors collect user information from various social networking sites and disambiguate users by taking into account the list of friends they share in the different social networks. These ideas can be easily implemented by an attacker who wants to match user profiles across different networks and construct a complete (aggregated) user profile.

A weak point that attackers exploit in order to connect information from different networks and reveal a user's identity are bugs in the browser monitoring mechanisms, such as cookies and browsing history. Malicious sites, with simple Javascript code can easily retrieve history information from popular browsers (history-sniffing) when a user visits them [7]. In their research, Wondracek et al. [34] used a similar history stealing technique (from the users' browsers) and managed to de-anonymize correctly 42 % (unique fingerprint) of the targeted users, when a social network user visits a malicious website.

23.3.4.5 Inference Attacks

Under the umbrella of identity exposure, we can also put the inference attacks. These attacks employ data mining techniques and data from within and outside the network in order to infer and reveal parts of a user identity (e.g., sex, age, habits, preferences, etc.). When sensitive data are published within the boundaries of the social network, or outside of it, the issue of privacy-preserving data publishing arises. For example, when a hospital publishes the patient records to a research center, it is necessary to protect patients' anonymity and guarantee that record linkage is not possible.

Fung et al. [12] provide a comprehensive survey of privacy-preserving data publishing techniques (k -anonymity being the most popular among them), which

can assist HSNs owners to protect the association of sensitive information with specific network members. Gkoulalas-Divanis et al. [13] present a survey of algorithms for publishing patient-specific data in a privacy-preserving way and identify three types of possible threats from inference attacks on anonymized published data: identity, membership and attribute disclosure.

23.3.4.6 Identity Theft

Through identity theft, an attacker can gain access to user accounts and profiles and consequently to their contacts and communications [37]. This can possibly harm the reputation or credibility of the real owner's profile while he/she is unaware of the attack. The attacker can simply pretend that he/she is the owner of the profile and use it to communicate with potential "victims" (e.g., pretend to be a doctor and give false advices to patients). The identity theft is rarely due to technical reasons and is mainly due to user ignorance about security precautions [5]. A similar threat to Identity Theft is Profile Cloning, where a user clones the profile and mimics another user in order to gain access to his/her social connections.

23.3.4.7 Phishing

The "phishing" is an act of deception, where the attacker impersonates a trusted entity, to acquire personal information such as sensitive private data and codes. Recently, such attacks on social networks have grown rapidly. According to Microsoft Security Intelligence an 84.5 % of all phishing attacks target social networks users. The huge amount of personal information available on social networks allows the perpetrator of such an attack to collect reliable information on his victims and then to cheat by sending them an e-mail, which appears to come from a legitimate user. Experiments conducted by the Indiana University showed that phishing attacks via e-mail and by collection of data from social networks have been successful in 72 % of the cases [16].

23.3.4.8 Communication Tracking

In the past, health care was managed mainly via interpersonal communication between the caregiver and the patient, while today, social media offers different modes of interaction. Recent studies in Internet-delivered therapy, especially for issues such as anxiety and depression show that on-line communication is more effective in alleviating mild to moderate symptoms than other methods of searching for health advice on-line [14]. This is expected to further increase the amount of doctor-patient communication data in the near future.

Communication tracking is a privacy threat that targets mainly the information that users exchange in their everyday communication within the network. By

monitoring users' communications the attacker manages to collect much more information than is available in their profile. This attack can also be performed by automatically traversing all user comments in a social network.

23.3.4.9 Information Leakage

Social network members exchange information with their friends and other network members and frequently voluntarily share sensitive data. In a survey with 166 participants, Torabi and Beznosov [29] observed that 95.8 % of users had shared some health data through their personal accounts on social networks. Such practices by users of social networks can result in the leakage of sensitive personal data.

Dossia, Microsoft HealthVault and Google Health in the past, provide personal health record management services, and allow users to store medical information, as well as personal information, on their central servers. Although they publish a privacy policy, these policies do not include confidentiality towards the provider and they do not allow patients to check whether the provider complies with their privacy policy. Information leakage incidents may occur due to different reasons:

- **Improper and incomplete privacy settings by the data owner.** In several cases, it becomes very complicated for the user to define his/her own privacy settings, and the predefined privacy policies set by the application provider are not adequate to cover all issues.
- **Unauthorised access by third-party applications.** Especially, when multiple applications in the cloud can gain access to a user profile or settings, there is always a risk for these applications to ask for more privileges than necessary, or to take advantage of their privileges for wrong purpose.

In order to conclude this long list of privacy threats for the users of social networks, we must add the results of an ENISA report on the top rising risks for social networks. According to this report the top risks comprise: malicious software, information leakage, phishing, spam, identity theft.

23.4 Privacy Requirements for HSNs

23.4.1 Privacy as System Requirement

In order to transform privacy from a general concept to a technical requirement included in information systems development, we need a set of privacy requirements [18]. According to the Common Criteria for Information Technology Security Evaluation standard (CC) these requirements are:

- **Anonymity:** The state of a subject being non-identifiable within a set of subjects (known as the *anonymity set*).

- **Pseudonymity:** The state of using a pseudonym as user ID.
- **Unlinkability:** Defined for two or more items, unlinkability assumes that within the system, these items are no more and no less related than they are related concerning the a-priori knowledge.
- **Unobservability:** The state of an item of interest being indistinguishable from any other item of interest.

With all the aforementioned requirements in mind, which are formally defined in [22], privacy can be defined as a set of technical requirements which prevent the disclosure of the identity of a user [35]. Additional requirements may comprise: authentication, authorization, identification, and data protection.

23.5 Enhancing Privacy in OSNs and HSNs

Privacy enhancing technologies comprise the tools, applications and devices designed to protect personal data, and assist Internet users in maintaining their privacy and anonymity. They can also be considered as a series of measures that protect privacy by eliminating or minimizing any unnecessary or undesirable processing of personal data, without loss of functionality of the information system [30]. In particular, privacy enhancing technologies provide the following options:

- They allow to minimize or eliminate disclosure, and collection of personal data or user identification data.
- They give users the possibility to carry out transactions without disclosing their identity.
- They allow users to exercise control over their personal data.
- They assist companies and organizations to implement policies and practices for privacy protection.

Privacy enhancing technologies have been categorised in the past using different classification schemes. The criteria for categorizing PETs can be the purpose they serve (e.g., minimization of disclosure and collection of personal data) or the technology they use to protect privacy (e.g., anonymization, encryption, etc.).

According to FIDIS (Future of Identity in the Information Society) project classification PET tools are divided to *transparency* and *opacity* tools [11]. The former aim at increasing users' awareness of the processes and practices that are followed when their personal data are processed and at helping them understand the potential consequences of data processing. Database audit interfaces, audit agents and log files are some examples of transparency tools.

Opacity tools, on the other hand, aim at concealing the identity of users or to prevent linkage between users and data. Some generic examples of tools that preserve identity privacy in social networks are MixMaster anonymous e-mail, TOR anonymizing web surfing, Pseudonyms, etc. Not all these tools are useful in HSNs, especially when anonymity is not an option and users can only use

pseudonyms. However, the awareness of network members about privacy threats and vulnerabilities is important and HSNs' owners (e.g., hospitals) work in this direction. For example, hospitals issue guidelines concerning how practitioners must use social networks, what information to disclose, etc.

The Meta Group survey⁵ performed on behalf of the Danish government divides PETs into two main categories. The first category concerns privacy protection and the second privacy management. More specifically, the first category includes tools and technologies directly involved in the protection of privacy by concealing information or eliminating the need for personalization of information, while the second includes tools that support the management of privacy rules.

According to the same survey, the purpose of PETs is to cover the four main requirements for privacy mentioned in Sect. 23.4 (i.e., unobservability, unlinkability, anonymity and pseudonymity). In addition, they provide secondary tools (e.g., tools for addressing spam, undesired Web content, or even unauthorized programs such as spyware, viruses) and informational tools, which assist users in understanding privacy issues. The long list of such tools comprises, amongst others: (a) *privacy protection tools*, such as pseudonymization tools, anonymization products and services, encryption tools, filters and blockers, track and evidence erasers, etc; and (b) *privacy management tools*, such as information and administrative tools.

In a classification provided by Koorn et al. [19] PETs are grouped into four categories: (1) *general*—includes encryption tools, logical access tools, biometrics, etc., (2) *data separation*—they separate the processing of data that identify a person from all other data to avoid cross linking, (3) *data anonymization* tools, such as MIX routers, Onion routers, cookie management tools, etc., and (4) *privacy management systems*—they automate the enforcement of security policies and compliance control tools, such as P3P (Platform for Privacy Preferences Project).

Finally, the technical report of Shen and Pearson (2011) classifies PETs into five categories as follows:

- PETs for anonymization.
- PETs to protect network invasion.
- PETs for identity management.
- PETs for data processing.
- Policy-checking PETs.

Privacy enhancing solutions comprise data filtering and minimization, anonymization, noise addition, etc. In metadata approaches, privacy policies are injected in the form of privacy rules that control access to data from different

⁵Privacy Enhancing Technologies—META Group Report v 1.1. 2005. Available for download from: <https://danskprivacynet.files.wordpress.com/2008/07/rapportvedrprivacyenhancingtechnologies.pdf>

applications. HL7 is working on a flexible standard that applies privacy and security labels to segments of users' personal data.⁶

According to a white-paper issued by the Healthcare Information and Management Systems Society (HIMSS) about Privacy and Security Considerations from the use of Social Media in Healthcare,⁷ privacy in health related social media and social networks is achieved through a multi-step approach, which comprises the following actions:

- Perform a social media risk assessment.
- Develop an overarching, risk-based, social media strategy consistent with organizational goals and objectives.
- Define a strategy to protect the organization's on-line reputation and brand from harm.
- Develop social media policies and procedures.
- Educate staff (e.g., doctors, nurses) and patients.
- Minimize regulatory and other legal/liability risks.
- Proactively monitor social media for compliance.

Yeratziotis et al. [36] claim that security and privacy do not come alone in healthcare social networks. According to the authors, it is vital that the development of security and privacy features for applications and websites are assessed for their usability, which will consequently increase the continuous and effective utilisation of the provided services. The authors propose a framework that consists of three components: (1) a three-phase process, (2) a validation tool, and (3) a usable security heuristic evaluation, as well as propose a list of items that must be checked to ensure usable security and privacy. This list comprises user awareness about security and privacy issues, user control on the privacy and security restrictions, and different levels of configuration details depending on the users' expertise.

In this line, the concept of Privacy-by-Design is gaining attention in the design of HSNs. It is a system engineering approach that was developed by the former Information and Privacy Commissioner of Ontario, Dr. Ann Cavoukian, back in the 90's. It aims to encourage system designers to take privacy into account throughout the whole engineering process.

Privacy-by-Design was initially expressed by deploying PETs. However, after several years of rather unsuccessful experimentation, it is clear that a more comprehensive approach is required. Privacy cannot and must not be an add-on, but it must be embedded into the design of the HSN.

⁶HL7 Privacy, Access and Security Services (PASS) Specification. Ann Arbor, MI, USA: HL 7 International. Available for download from: <http://wiki.siframework.org/file/view/PASS+Access+Control+Conceptual+Model+Release+1.0.pdf>.

⁷Social Media in Healthcare: Privacy and Security Considerations. HIMSS White Paper. Available for download from: http://himss.files.cms-plus.com/HIMSSorg/Content/files/Social_Media_Healthcare_WP_Final.pdf.

Ann Cavoukian has identified seven Foundational Principles that must be practised in order to achieve the Privacy-by-Design objectives, and that the architects and operators of HSNs must follow. Namely:

1. **Proactive not Reactive; Preventative not Remedial:** Privacy-by-Design anticipates and prevents privacy-invasive events before they happen. It aims to prevent them from occurring.
2. **Privacy as the Default Setting:** Privacy must be the default state of the system. No action must be required on the part of the individual.
3. **Privacy Embedded into Design:** Privacy is not an add-on. It must be embedded into the design and architecture of IT systems.
4. **Full Functionality—Positive-Sum, not Zero-Sum:** Privacy-by-Design avoids trade-offs between design goals. For instance, it demonstrates that both privacy and security are possible.
5. **End-to-End Security—Full Life-cycle Protection:** Privacy-by-Design seeks to protect data throughout their entire life-cycle, i.e., at the end of the process, all data are securely destroyed.
6. **Visibility and Transparency—Keep it Open:** Privacy-by-Design gives the opportunity to verify the efficiency of the privacy protection mechanism through transparency.
7. **Respect for User Privacy—Keep it User-Centric:** Privacy-by-Design requires the system designers to keep the interests of the user uppermost by empowering user-friendly options.

23.6 On-line Social Networks in the Healthcare Domain

Healthcare social networks and user communities can be build on top of existing social networking applications or on new platforms, which are designed on purpose. The survey work of Grajales et al. [15] provides a good summary of social networking applications that host healthcare networks. More specifically, authors analyse different categories of social media which have been used by healthcare professionals, patients and researchers, such as: (1) blogs (e.g., WordPress), (2) microblogs (e.g., Twitter), (3) social networking sites (e.g., Facebook), (4) professional networking sites (e.g., LinkedIn), (5) wikis (e.g., Wikipedia), (6) collaborative filtering sites (e.g., Digg), (7) media sharing sites (e.g., YouTube, Slideshare), and (8) 3-D virtual worlds (e.g., SecondLife). They also examine thematic networking sites (e.g., 23andMe) and application mashups (e.g., HealthMap).

Healthcare requires privacy and anonymity, making traditional social media sites such as Facebook and Twitter inadequate in sharing health data or personal health experiences. This is where platforms designed specifically for supporting health related social networks come in to the scene.

23.6.1 Advice Seeking Networks

Patients, on the other side, access specialized social networks which focus on healthcare subjects—especially on specific diseases, research and support around them. In such networks, they are encouraged to connect with other patients, share their stories, and get informed about their disease.

CureDiva,⁸ is a social network and on-line e-shop, targeting breast cancer patients. It has a privacy mechanism, which allows members to choose their preferred level of privacy on personal content. The network creates a personalized experience for its members in order to increase their comfort.

MedHelp⁹ is a social media site that aims in consumer health engagement. It has 13 million active monthly users, over 200 condition-specific communities and provides expert forums for the health professionals to answer the questions from health consumers directly. It offers community forums and expert blogs and a data platform for consumers to gather data from mobile apps, web apps, and health devices.

E-couch¹⁰ is an on-line communication tool, which provides several self-help interactive programs that allows patients with depression, generalised anxiety and worry, social anxiety, relationship breakdown, and loss and grief to contact experts and ask for their support.

23.6.2 Patient Communities

ConnectedLiving¹¹ is a private social network, which interconnects residents of nursing homes, assisted living complexes, and other senior housing centres and aims in creating a community of seniors, which are currently the most disconnected part of the population. Network-members form friendship bonds with their real-world friends within the social network and in the same time can grant access to external social networks contents such as the members' profile in Instagram or Facebook.

The popularity of smart-phones and the emerge of mobile health gave rise to data intensive applications and networks where patients perform self-tracking and share their data with doctors and the community. Self-trackers are using applications to monitor sleep, food intake, exercise, blood sugar and other physiological states and behaviours. Patients use the data in order to receive alerts for their health, physicians suggest these solutions to patients so that they can have real-time feedback on the results of a treatment and be able to adjust therapies faster [28].

⁸<https://www.curediva.com/>.

⁹<http://MedHelp.org>.

¹⁰<https://ecouch.anu.edu.au>.

¹¹<http://www.connectedliving.com/>.

PatientsLikeMe,¹² Smart Patients,¹³ FaceToFaceHealth¹⁴ are a few out of many on-line social networks where patients share their experience using patient-reported outcomes, find other patients like them matched on demographic and clinical characteristics, and learn from others to improve their way of living. The goal of such websites is to help patients answer the question: “Given my status, what is the best outcome I can hope to achieve, and how do I get there?” [33].

23.6.3 Professional Networks

Sermo¹⁵ is the most popular social network for doctors, which however, limits its membership to US-based doctors. NurseTogether¹⁶ is a similar professional networking community for nurses. Doximity¹⁷ is a professional social network for doctors only, founded in 2010, which resembles more to LinkedIn than Facebook. In contrast to Sermo that allows anonymous postings, Doximity insists on real names. Membership is validated by DEA number, which guarantees the true identity of a member. The two networks have very different styles, since Sermo focuses more on discussion forums, while Doximity emphasizes professional networking and private messaging. In this direction, Doximity members have the ability to selectively share contact information (e.g., cellphone number) with other members.

Each network type poses different challenges in terms of meeting the privacy requirements. A comprehensive study of the controls needed in order to achieve each requirement for such network types must follow in order to ensure privacy for all types of networks. One of the main challenges that need to be further addressed is the conflict between legal restrictions, human privacy restrictions and the need for immediate access to a patient’s data when his health is in danger.

23.7 Conclusion

The use of Social networks and social media from patients and doctors, for health related issues hides several privacy threats and risks, which must be properly addressed due to the sensitive nature of patient information. Indicative types of social networks for health include professional networks, advice seeking applica-

¹²<https://www.patientslikeme.com/>.

¹³<https://www.smartpatients.com/>.

¹⁴<http://www.facetofacehealth.com>.

¹⁵<http://sermo.com>.

¹⁶<http://www.nursetogether.com>.

¹⁷<https://www.doximity.com>.

tions and patient communities. This work, presented privacy risks, requirements and available solutions, and made a first step towards a best practice guide, which will outline both the technical and procedural countermeasures required in order to maintain privacy, taking into account modern technology environments.

References

1. Apostolakis, I., Koulierakis, G., Berler, A., Chryssanthou, A., Varlamis, I.: Use of social media by healthcare professionals in greece: an exploratory study. *Int. J. Electron. Healthc.* **7**(2), 105–124 (2012)
2. Barnes, J.A.: Class and Committees in a Norwegian Island Parish. Plenum, New York (1954)
3. Bennett, C.J., Parsons, C., Molnar, A.: Forgetting, non-forgetting and quasi-forgetting in social networking: Canadian policy and corporate practice. In: *Reloading Data Protection*, pp. 41–59. Springer, New York (2014)
4. Chryssanthou, A., Varlamis, I., Latsiou, C.: A risk management model for securing virtual healthcare communities. *Int. J. Electron. Healthc.* **6**(2–4), 95–116 (2011)
5. Cutillo, L.A., Manulis, M., Strufe, T.: Security and privacy in online social networks. In: *Handbook of Social Network Technologies and Applications*, pp. 497–522. Springer, New York (2010)
6. Damschroder, L.J., Pritts, J.L., Neblo, M.A., Kalarickal, R.J., Creswell, J.W., Hayward, R.A.: Patients, privacy and trust: patients' willingness to allow researchers to access their medical records. *Soc. Sci. Med.* **64**(1), 223–235 (2007)
7. Eckersley, P.: How unique is your web browser? In: *Privacy Enhancing Technologies*, pp. 1–18. Springer, New York (2010)
8. Ellison, N.B., et al.: Social network sites: Definition, history, and scholarship. *J. Comput.-Mediat. Commun.* **13**(1), 210–230 (2007)
9. Eysenbach, G.: Infodemiology and infoveillance: tracking online health information and cyberbehavior for public health. *Am. J. Prev. Med.* **40**(5), S154–S158 (2011)
10. Finn, R.L., Wright, D., Friedewald, M.: Seven types of privacy. In: *European Data Protection: Coming of Age*, pp. 3–32. Springer, New York (2013)
11. Fritsch, L.: State of the art of privacy-enhancing technology (pet). Deliverable D2 **1** (2007)
12. Fung, B., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: a survey of recent developments. *ACM Comput. Surv.* **42**(4), 14 (2010)
13. Gkoulalas-Divanis, A., Loukides, G., Sun, J.: Publishing data from electronic health records while preserving privacy: a survey of algorithms. *J. Biomed. Inform.* **50**, 4–19 (2014)
14. Glozier, N., Christensen, H., Naismith, S., Cockayne, N., Donkin, L., Neal, B., Mackinnon, A., Hickie, I.: Internet-delivered cognitive behavioural therapy for adults with mild to moderate depression and high cardiovascular disease risks: a randomised attention-controlled trial. *PLoS ONE* **8**(3) (2013)
15. Grajales III, F.J., Sheps, S., Ho, K., Novak-Lauscher, H., Eysenbach, G.: Social media: a review and tutorial of applications in medicine and health care. *J. Med. Internet Res.* **16**(2), e13 (2014)
16. Hogben, G.: Security issues and recommendations for online social networks. ENISA Position Paper **1**, 1–36 (2007)
17. Holvast, J.: History of privacy. In: Matyas, V., Fischer-Hubner, S., Cvrcek, D., Svenda, P. (eds.) *The Future of Identity in the Information Society. IFIP Advances in Information and Communication Technology*, vol. 298, pp. 13–42. Springer, Heidelberg (2009). doi:10.1007/978-3-642-03315-5_2. http://dx.doi.org/10.1007/978-3-642-03315-5_2
18. Kalloniatis, C., Mouratidis, H., Vassilis, M., Islam, S., Gritzalis, S., Kavakli, E.: Towards the design of secure and privacy-oriented information systems in the cloud: identifying the major concepts. *Comput. Stand. Interfaces* **36**(4), 759–775 (2014)

19. Koorn, R., van Gils, H., ter Hart, J., Overbeek, P., Tellegen, R., Borking, J.: Privacy enhancing technologies, white paper for decision makers. Ministry of the Interior and Kingdom Relations, the Netherlands (2004)
20. Krishnamurthy, B., Wills, C.E.: On the leakage of personally identifiable information via online social networks. In: Proceedings of the 2nd ACM Workshop on Online Social Networks, pp. 7–12. ACM, New York (2009)
21. Mitrou, L.: Privacy challenges and perspectives in europe. In: An Information Law for the 21st Century, pp. 704–718. Nomiki Vivliothiki, Athens (2011)
22. Pfitzmann, A., Hansen, M.: A terminology for talking about privacy by data minimization: Anonymity, unlinkability, unobservability, pseudonymity, and identity management-version v0. 34. TU Dresden and ULD Kiel, Tech. Rep (2011)
23. Robinson, N., Graux, H., Botterman, M., Valeri, L.: Review of eu data protection directive: summary. Information Commissioner's Office (2009)
24. Rosenberg, R.S.: The Social Impact of Computers. Elsevier, Amsterdam (2013)
25. Rowe, M., Ciravegna, F.: Disambiguating identity through social circles and social data. In: 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008) (2008). Proceedings of the CISWeb Workshop, located at the 5th European Semantic Web Conference ESWC 2008 Tenerife, Spain, 2 June 2008
26. Serenko, N., Fan, L.: Patients' perceptions of privacy and their outcomes in healthcare. *Int. J. Behav. Healthc. Res.* **4**(2), 101–122 (2013)
27. Stuart, A.H.: Online privacy policies: contracting away control over personal information? *Penn State Law Rev.* **111**(3), 587–624 (2007)
28. Swan, M.: Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. *Int. J. Environ. Res. Public Health* **6**(2), 492–525 (2009)
29. Torabi, S., Beznosov, K.: Privacy aspects of health related information sharing in online social networks. In: Proceedings of the 2013 USENIX Conference on Safety, Security, Privacy and Interoperability of Health Information Technologies, pp. 3–3. USENIX Association, Berkeley (2013)
30. Van Blarkom, G., Borking, J., Olk, J.: Handbook of privacy and privacy-enhancing technologies. Privacy Incorporated Software Agent (PISA) Consortium, The Hague (2003)
31. Warren, S.D., Brandeis, L.D.: The right to privacy. *Harv. Law Rev.* **4**, 193–220 (1890)
32. Westin, A.F.: Privacy and freedom. *Wash. Lee Law Rev.* **25**(1), 166 (1968)
33. Wicks, P., Massagli, M., Frost, J., Brownstein, C., Okun, S., Vaughan, T., Bradley, R., Heywood, J.: Sharing health data for better outcomes on patientslikeme. *J. Med. Internet Res.* **12**(2), e19 (2010)
34. Wondracek, G., Holz, T., Kirda, E., Kruegel, C.: A practical attack to de-anonymize social network users. In: 2010 IEEE Symposium on Security and Privacy (SP), pp. 223–238. IEEE (2010)
35. Yanes, A.: Privacy and anonymity. arXiv preprint (2014) [arXiv:1407.0423]
36. Yeratziotis, A., Van Greunen, D., Pottas, D.: A framework for evaluating usable security: the case of online health social networks. In: HAISA, pp. 97–107 (2012)
37. Zilpelwar, R.A., Bedi, R.K., Wadhai, V.: An overview of privacy and security in SNS. *International Journal of P2P Network Trends and Technology*, **2**(1), 23–26 (2012)

Part IV

**Privacy Through Policy, Data
De-identification, and Data Governance**

Chapter 24

Privacy Law, Data Sharing Policies, and Medical Data: A Comparative Perspective

Edward S. Dove and Mark Phillips

Abstract The sharing and linking of medical data across borders is now a key enabler of new medical discoveries. Data are no longer simply collected and used at a single physical site, such as a laboratory or a research institute. Instead, communication flows between research teams within and across national borders bring together the necessary data and expertise to clarify previously unknown disease aetiologies. Integration of medical data and secure health records systems now allows clinicians to develop early treatment strategies tailored to a specific patient. As policymakers, patient advocacy groups, and biomedical researchers gravitate toward recognizing the benefits of global data sharing, they may be challenged by regulatory systems that were developed when the norm was using and sharing medical data only within a single jurisdiction. This chapter describes and compares key data privacy legal frameworks (Canada, US, UK, EU, Council of Europe, OECD) and discusses data sharing policies adopted by major biomedical research funding organisations (the NIH, Canadian Institutes of Health Research, Genome Canada, Wellcome Trust) in the context of their impact on medical data privacy. In so doing, the chapter explains not only the content, significance, and practical usefulness of these laws, regulations, and policies as they relate to medical data, but also identifies lingering barriers to global data sharing and suggests ways to overcome them while maintaining robust data privacy protection.

24.1 Introduction

Data, including medical data, are the currency of the twenty first century. In our information society, using and sharing medical data are critical to achieving both translational and precision medicine, such as improved disease classification based

E.S. Dove (✉)

School of Law, University of Edinburgh, Edinburgh, UK

e-mail: edward.dove@ed.ac.uk

M. Phillips

Centre of Genomics and Policy, McGill University, Montreal, H3A 0G1, Canada

e-mail: mark.phillips2@mail.mcgill.ca

on molecular profiles allowing tailored treatments, interventions, and models for prevention [1, 17]. Medical data are now employed regularly “to support evidence-based decision-making, to improve the quality of care provided, and to identify and achieve cost efficiencies” [15].

Medical data are used not only to deliver necessary healthcare directly to individuals—secondary uses of medical data to broaden scientific knowledge, for both public and private benefit, are also myriad and increasing. Data have long been used for invaluable secondary purposes that benefit society as a whole, such as population health monitoring, healthcare quality improvement, and biomedical research. Moreover, they are no longer collected and used only within single sites such as clinics, laboratories, or research institutes; communication flow within and across national borders and research teams, encompassing data from clinical and population research, enables researchers and clinicians to connect the diverse types of datasets and expertise needed to elucidate the molecular basis and complexities of disease aetiology. The number of large-scale health research projects that involve the collection of whole genome sequencing data is continuing to rise in countries around the world.

This integration of medical data has begun to allow explanations of the aetiologies of cancer, inherited diseases, infectious diseases, and drug responses. Furthermore, integrating medical data with electronic medical or health record systems, securely stored in research and clinical databases, can help clinicians to develop earlier and more targeted treatment strategies for their patients. The growth of “eHealth” technologies is streamlining the flow of medical data within and across borders to improve quality of care and service delivery, and reduce healthcare system inefficiencies and costs.

Using and sharing medical data, however, *requires regulatory systems that protect privacy*. Depending on the context in which they are used and how they are related to other information, medical data, whether processed by healthcare providers in the delivery of healthcare or by researchers in the furtherance of generalizable knowledge, can be highly sensitive. Respondents to a survey commissioned by the Wellcome Trust in 2013, for example, strongly felt “that personal medical data are confidential, private and sensitive, and should not be shared outside secure, authorised bodies . . . and especially not with private companies such as employers, insurance providers, and drug manufacturers. Mental health data was sometimes regarded as particularly personal and sensitive” [77]. Studies have shown “that most adults are concerned about the security and privacy of their [medical] data, and such concerns are associated with an increased likelihood of non-disclosure of sensitive information to a healthcare professional” [2].

On the other hand, in some contexts medical data may not be so sensitive. In a 2013 survey prepared for the Office of the Privacy Commissioner of Canada, for example, “Canadians were asked, in an open-ended manner, what risks to their own privacy concerned them the most . . . Financial information/bank fraud topped the list, with nearly a quarter citing it (23%); medical data ranked near the bottom at 3 % [58]. Though medical data are not invariably more sensitive than other types of data, it is certain enough that data processed in a medical context touch on

important personal privacy interests. Therefore, medical data relating to individuals (and arguably, families and communities as well) must be safeguarded, and their sharing controlled through appropriate legislation and policy, including sound ethics review oversight and data access mechanisms.

But what are medical data exactly? Some legal and policy instruments draw distinctions between the terms “medical data”, “health data”, “health information”, “protected health information”, “personal health information”, and “data relating to health”. For example, the Article 29 Data Protecting Working Party [4–6] (an independent European advisory body on data protection and privacy set up under Article 29 of Europe’s Data Protection Directive 95/46/EC [28]) interprets “health data” as a much broader term than “medical data”. In particular, they have opined that everything from the wearing of glasses or contact lenses, to membership in a support group with a health-related objective, to data about a person’s intellectual and emotional capacity, or smoking and drinking habits, constitutes health data. Medical data, on the other hand, are a narrower category of data comprising the physical or mental health status of a “data subject” that are generated in a professional, medical context. As the Working Party explains:

This includes all data related to contacts with individuals and their diagnosis and/or treatment by (professional) providers of health services, and any related information on diseases, disabilities, medical history and clinical treatment. This also includes any data generated by devices or apps, which are used in this context, irrespective of whether the devices are considered as “medical devices” [6].

This narrow position contrasts with the Council of Europe’s Recommendation No. R (97) 5 on the Protection of Medical Data, which defines medical data as “all personal data concerning the health of an individual [and those] data which have a clear and close link with health as well as . . . genetic data” [19]. Despite these definitional nuances, each approach speaks to data about an identifiable individual that are related to the individual’s health or the provision of health services to the individual. At the same time, these data touch on morally relevant values and interests that transcend the individual [55]. Importantly, they reveal intimate aspects, especially in the case of genomic data, of family members. Medical data encompass information about lifestyle and behaviour; health conditions and concerns; history of healthcare procedures and medication use; results of medical tests; related information about family members and other individuals; and genetic information about individuals and their blood relatives. Misuse of medical data, especially their unwarranted disclosure, “could adversely affect the opportunities available to individuals, including eligibility for loans, healthcare, employment and educational opportunities, or adoption” [7]. Misuse could also lead to serious repercussions for researchers, healthcare practitioners, and their organisations. Experience demonstrates that individuals will only share their medical data, or participate in or trust healthcare systems and research studies, if they know that their data are sufficiently protected. Robust privacy protection is therefore needed, and it is the responsibility of governments, organisations, and individuals to effectively protect medical data.

Yet while medical data sharing and collaboration are increasingly embraced by policymakers, funders, patient advocacy groups, and the international biomedical research community, inefficiencies and insufficiencies remain, in part due to generation-old data privacy regimes originally developed to protect personal data within single jurisdictions [63]. These regimes are not attuned sufficiently to the evolving paradigm of large-scale, global, and data-driven biomedical research, and thus often result in inefficient data flow and unnecessary data transfer costs and delays. It may well be time to rethink how data privacy regimes are conceived, and how they can both promote medical data sharing and protect medical data privacy in a proportionate manner.

In this chapter, we provide an overview and international comparison of data privacy laws and regulations in several key jurisdictions (including Canada, the United States, and Europe) and also discuss data sharing policies in the context of medical data privacy. We intend to illustrate not only the content, significance, and practical usefulness of these laws, regulations, and policies as they relate to medical data, but also to identify lingering undue regulatory barriers to data sharing and suggest ways to overcome them while maintaining robust and proportionate privacy safeguards.

24.2 Overview of Data Privacy Legal Frameworks

Privacy is considered to be a fundamental interest in almost all societies and a fundamental right in some, such as in Europe. It is an ancient concept that has been discussed in foundational philosophical and legal treatises (e.g., Aristotle's *Politics* from approximately 350 B.C., John Locke's *Second Treatise of Government* from 1690), and has been implemented in law for thousands of years [22]. For example, the notion of the “private sphere”, as equating to the interests of individuals, as distinct from a “public sphere”, relating to political activities, was incorporated into late Roman law, including in the first chapter of the two sections of the *Corpus Juris Civilis*, the compilation of Roman law issued by Emperor Justinian in 533–534 CE [60].

Yet privacy is a notoriously vague term and is contextual [54]. As one report cautions, “discussion of privacy across nations and cultures must be sensitive to the impact of cultural norms and environments before applying universal concepts of privacy” [78]. While taking note of numerous unresolved philosophical, ontological, and semantic debates, we define privacy, at least in its informational dimension, as *a state of affairs whereby data relating to a person are either in a state of non-access, or in a state of managed access such that the person is able to decide whether and how they may be used and shared, and to know how those data are actually used and shared*. Privacy is instrumentally valuable, as it enables people to flourish in developing personal relationships and social participation, and it is intrinsically valuable, as it is grounded in the values of dignity, integrity, and autonomy. This

consensus is reflected, albeit to varying degrees, in human rights declarations as well as in data privacy legislation and policies across much of the world.

The number of global data privacy regulations and policies is vast. There are at least 109 countries in the world as of 2015 with data privacy laws in force [35]. Despite this regulatory expanse, all data privacy legal frameworks, whether specific to medical or other types of data, seek to create a realm of legal certainty in which the processing of personal data can occur in a way that both benefits society and protects individuals from harm.

Two observations deserve mention at the outset. First, data privacy legal frameworks tend to address the “processing” of personal data, which is generally understood to include any operation or set of operations performed on personal data and therefore includes: collection; recording; organisation; structuring; storage; adaptation or alteration; retrieval; consultation; use; disclosure by transmission, dissemination, or otherwise making available; alignment or combination; and erasure. Second, data privacy legal frameworks usually apply only to data that relate to an identified or identifiable individual person, known as the “data subject” (hence the terms “personal data” or “personal information”). This traditional definition is challenged in the genetic context by the relational nature of genetic data [62].

Additionally, two caveats are in order. First, though the two concepts are closely linked, privacy is not synonymous with data protection. Privacy is a broader concept that embodies a range of rights and values, such as the right to be let alone, intimacy, seclusion, and personhood. It can include control over personal data, but not all personal data are private. Moreover, control over personal data might rest with a person, but this is not a necessary condition for privacy. If it were, it would mean that every loss of control was a loss of privacy and we know this is not so. Data protection, on the other hand, seeks to protect values other than just privacy, such as data security, data quality, non-discrimination, and proportionality. Data protection is a “set of legal rules that aims to protect the rights, freedoms, and interests of individuals, whose personal data are collected, stored, processed, disseminated, destroyed, etc”. The ultimate objective is to ensure “fairness in the processing of data and, to some extent, fairness in the outcomes of such processing.” The fairness of processing is safeguarded by a number of principles [64].

Furthermore, “unlike privacy’s elusive and subjective nature that makes the right different in different contexts and jurisdictions, data protection has an essential procedural nature that it makes it more objective as a right in different contexts” [64]. Because some jurisdictions refer to “privacy laws” and others refer to “data protection laws”, though both speak to protection of personal data, in this chapter we use the imperfect but relatively catch-all term “data privacy laws”.

Second, we note that privacy is distinct from confidentiality, which relates instead to the protection of information that a person has disclosed in a relationship of trust with the expectation that this information will not be divulged to others without permission or authorisation. The duty of confidentiality is reflected in self-regulatory codes of medical and research professionals, such as codes of ethics and medical data privacy codes, as well as in law such as the common law duty of confidentiality, not to mention customary practices that have evolved in trusting relationships.

In law, data privacy frameworks are seen as protecting the personal data component of privacy. Comprehensive human rights laws and constitutions often address some element of personal data or informational privacy, classically framed as protection of individuals' privacy interests against arbitrary state interference. For example, in Canada, informational privacy is protected under the country's constitution (specifically in the Canadian Charter of Rights and Freedoms) as an implicit component of its guarantees of liberty, especially the protection against unreasonable search or seizure by state actors.

Data privacy laws as such emerged relatively recently. Legislation protecting certain forms of data was first enacted in 1970 by the German state of Hesse (the statute is known as the *Hessisches Datenschutzgesetz*), followed by Sweden in 1973 (the *Datalagen*) and, subsequently, by other countries, both in Europe and elsewhere (e.g., the US Privacy Act of 1974, which notably applies only to personal data held in record systems of federal agencies). Comprehensive "omnibus" data privacy legislation emerged in the mid-1980s. Modern data privacy principles are often traced to three reports from the 1970s [10, 47]: one written by a committee created by the British Parliament to investigate privacy problems related to the use of computerized data records in the commercial sector; another by an advisory committee to the US Department of Health, Education and Welfare (HEW), which called for "attention to issues of record-keeping practice in the computer age that may have profound significance for us all" [37]; and another from a study commission by the US Congress that called for the protection of privacy in several sectors, including healthcare and research [73]. The British parliamentary committee drafted a set of regulatory principles [79] quite similar to the safeguards recommended by the HEW report, which detailed fundamental principles of "fair information practice" that have since been adopted in various forms at the international level, as discussed in the following section. The international emergence of data privacy laws in the 1970s and 1980s aimed to address privacy issues generated by new technologies such as vastly expanded computing, centralized processing of personal data, and the establishment of large data banks.

In the biomedical context, data privacy laws have two key objectives. First, they aim to enable free flows of data necessary for the delivery of health services, management of public health, and biomedical research. Second, and at the same time, they seek to establish appropriate privacy and security frameworks that protect personal data from misuse.

Much of the discussion surrounding data privacy can be described in terms of the appropriate relationship between these two objectives. The second objective can be seen either more as instrumentally valuable as a means of accomplishing the first, or as inherently valuable in itself. Where the two objectives conflict with one another, one will sometimes be argued to override the other in particular circumstances. There is consensus, however, that both objectives are worth fulfilling to the degree they are reconcilable.

Because of this split purpose, data privacy laws tend to have a wide reach, and over time their scope has expanded. To take one example, the Council of Europe's Recommendation No. R (97) 5 on the Protection of Medical Data provides

that legislation protecting data privacy should apply to all medical data, whether processed by a physician or by another person [19]. Also, the form in which personal data are stored or used is generally irrelevant to the applicability of data privacy laws; hence, cell samples of human tissue may be considered personal data as they record the DNA of a person, though this remains unsettled in many data privacy laws [11, 62].

Regrettably for biomedical research projects with an international character, globally harmonisation data privacy standards do not yet exist, though a concerted attempt was made with the Madrid Resolution of 2009 (International Conference of Data Protection and Privacy Commissioners 2009 [41]). Instead, data privacy frameworks generally fall within one of three categories: (1) comprehensive laws, meaning omnibus data privacy statutes often grounded in human rights principles; (2) sectoral laws that apply only to the demands of data privacy in a specific sector, such as healthcare (though few countries have adopted specific medical data privacy legislation); and (3) other rules or sets of rules that do not have the force of law, but that may nonetheless entail serious consequences when violated, such as professional codes of conduct or policy guidelines. Within each category there is tremendous variation.

Approaches to data privacy protection differ across world regions. As Professor Rolf Weber observes: “From a comparative perspective, European regulations are quite advanced, setting a high level of data protection. In the United States and in Asia, the emphasis is more on self-regulatory approaches” [75]. Some legal regimes operate on the scale of a regional grouping of countries, such as Europe; many apply only within a single country, while others are narrower still, for example, laws specific only to a single province or state within a country, or even to a specific type of industry. Laws may apply separately to public, private, or health institutions. Yet whatever their geographic, sectoral, or institutional scope, any data privacy legal framework may have a bearing on the processing of medical data. Consequently, the lack of globally harmonisation data privacy standards creates numerous risks. Professor Weber lists several (see Fig. 24.1) [75].

1. Non-compliance with national law.
2. Unauthorised release of personal information.
3. Inability to provide individuals with access to their personal information.
4. Inability to cooperate with national regulators in case of complaints.
5. Inability of the national regulator to investigate or enforce the law.
6. Inability to guarantee the protection of personal data in countries with a low protection level.
7. Conflicts between national and foreign laws.
8. Possible access to data by foreign governments.
9. Overseas judicial decisions requiring the disclosure of data.
10. Problems with recovery or secure disposal of data.
11. Loss of trust/confidence if data are transferred and misused.

Fig. 24.1 Risks created by the lack of globally harmonisation data privacy standards

Despite this variation, there are points of convergence. For instance, data privacy frameworks often impose specific duties on an individual or organisation who must determine the purposes and means of the personal data processing—the “data controller”, also known by other terms such as “information custodian”, “covered entity”, or “trustee”—as well as on an individual or organisation—known as the “data processor”—who processes personal data on behalf of the data controller. Another commonality is that generally, no restrictions apply to the processing of personal data where the person to whom they relate has specifically consented to that particular use, though the limitations of specific consent in the context of certain databases (e.g., cancer registries), infrastructures (e.g., biobanks), or large-scale or longitudinal biomedical research studies have been well documented (e.g., O’Neill [57]).

Privacy concerns may be assuaged by *anonymization* (also referred to as *de-identifying*) or *pseudonymising* (also referred to as *key-coding*) medical data before putting them to downstream uses [5]. To the benefit of researchers and medical data custodians, fewer legal restrictions may apply to data that have been anonymised or pseudonymised past certain thresholds. This said, many data privacy laws still consider pseudonymised data to be personal data, and the processing of personal data for the purposes of achieving anonymisation (i.e., the rendering of personal data to an anonymised state) is subject to data privacy laws because prior to such anonymisation, the data are still personal. But while scientific researchers tend to use anonymised data wherever possible, there are three main limits to anonymisation (see Fig. 24.2). Pseudonymisation largely lacks these weaknesses, and is a much simpler process, yet the trade-off is that it is a weaker privacy protection measure, and makes no claim to irreversibly obfuscating data subjects’ identities.

The growth of international biomedical research collaboration has led to concomitant exponential growth in cross-border data flows. Data privacy legal frameworks tend to adopt one of two general approaches with respect to cross-border “data transfer” (i.e., sharing data within and across jurisdictional borders). One set of frameworks requires that before medical data may be transferred, the data controller must take specific steps to ensure that the entity receiving the data is governed by “adequate” legislative oversight to protect the data (termed the *adequacy* approach). The other set requires that data controllers take what they consider appropriate steps to ensure adequate data protection during and after the transfer, holding the controller accountable for any improper use that may ultimately be made of the data by the transferee (termed the *accountability* approach). Although some laws incorporate elements of both approaches, the adequacy approach tends to be geographically oriented and relevant only to international data transfer: legislation in the country of the transferee will either be deemed to provide adequate protection, or not. The accountability approach, on the other hand, tends to be organisational and to apply even to transfer within the borders of a given country. It requires that the data transferee agree to be subject to sufficient legally binding privacy obligations, such as through a contract.

Aside from these general trends, each data privacy framework emerges within its own context, and each contains its own idiosyncrasies. As a comprehensive

1. The methods and degree of de-identification required to warrant fewer legal restrictions are not only inconsistent but almost always also unspecified, causing legal uncertainty when it comes to working with medical data [44]. This situation is partly a result of data privacy legislation that was rarely drafted with the specific issues of medical data privacy and contemporary biomedical research in mind. Another factor is that de-identification (and re-identification) is a rapidly developing – and also controversial – field, which makes it difficult to authoritatively lay down a specific and prospective standard for de-identification in law.
2. Data privacy research is now making clear that although a dataset may be anonymised according to conventional approaches (i.e., through randomization or generalization by considering publicly available datasets), its cross-linking with data available elsewhere (e.g., from another dataset) can make it possible to infer data subjects' identities. Large datasets, particularly those including extensive genomic information, cannot be completely safe from inferential exploitation and ultimately data subject re-identification [30]. Data controllers must therefore take appropriate steps to ensure that the medical data they hold are used only for acceptable purposes within the scientific scope of a research study and in accordance with the consent provided by patients or participants. Effort involved in rigorously anonymising data in this case is of limited (perhaps even of questionable) benefit.
3. Nor is anonymisation of medical data particularly beneficial when researchers or clinicians want to link medical data to other data sources over time (as anonymised data have been irreversibly de-linked); to re-contact donors to enhance research aims; to obtain additional information or invite participation in other research projects; to communicate a clinically actionable finding; to identify and correct errors or to amend medical data when new information becomes available; or to allow donors to withdraw their medical data from a study. Thus, while anonymisation may be championed as a means of achieving strong data privacy protection, in the medical data context it usually offers only limited utility to (not to mention respect for) both researchers and patient-participants alike. As a recent Nuffield Council on Bioethics report notes, “Faced with contemporary data science and the richness of the data environment, protection of privacy cannot reliably be secured merely by anonymisation of data or by using data in accordance with the consent from ‘data subjects’. Effective governance of the use of data is indispensable” [55].

Fig. 24.2 Three main limitations to anonymisation of personal data

international review is beyond the scope of this chapter, we now turn to discuss in a comparative manner some of the most influential data privacy laws, guidelines, and policies. The details of specific national data privacy laws and regulations are clearly set out in the leading textbooks (e.g., Beyleveld et al. [8]; Boniface [9]; Bygrave [10]; Greenleaf [34]; Kenyon and Richardson [42]; Kuner [45]; Power [59]; Solove and Schwartz [61]) and are available in the free access International Privacy Law Library (worldlii.org/int/special/privacy), located on the World Legal Information Institute (WorldLII) website.

24.3 Data Privacy Laws and Guidelines

International instruments have addressed data privacy in some form for nearly 50 years. As Bygrave [10] notes, as early as 1968, the United Nations General Assembly passed a resolution inviting the UN Secretary General to examine individuals' right to privacy "in the light of advances in recording and other techniques" [67]. A UN report in 1976 followed, calling for countries to adopt data privacy legislation and listing a set of minimum legislative standards [68]. In 1990, the UN General Assembly adopted a set of data privacy guidelines that set out minimum guarantees to be legislated nationally and to be respected by governmental international organisations to ensure responsible, fair, privacy-friendly data processing [69]. An important early effort to codify the principles of data privacy, including the international transfer of data, was made in 1980 by the Organisation for Economic Co-operation and Development (OECD) [56].

Yet while the OECD and other international organisations have promoted privacy as a fundamental value and a condition for the free flow of personal data across borders, as well as in inspiring data privacy legislation around the world, still, almost 50 years later, no set of principles or rules has been accepted as an authoritative global standard. Consequently, a variety of data privacy frameworks have proliferated.

24.3.1 *The OECD Privacy Guidelines*

Although the 1980 OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (OECD Privacy Guidelines [56]) are non-binding and not directly enforceable, even within OECD member countries, it would be difficult to overstate their influence. Nearly all existing privacy and data sharing laws and policies have adopted at least some of the principles they articulate. Turkey is the only OECD member country (of 34, currently), other than the US in relation to the private sector, which does not have a data privacy law implementing the OECD Privacy Guidelines [33].

The OECD Privacy Guidelines present minimum recommended standards that each OECD member country is encouraged to adopt for both public and private sectors, and to supplement according to its individual needs through the creation of domestic privacy and data transfer laws. Recently updated in 2013, the Privacy Guidelines allow considerable flexibility in implementation by data controllers and OECD member countries.

Their core consists of eight data privacy principles that have remained wholly intact following the 2013 revision (see Fig. 24.3). These principles, as well as the broader guidelines, apply only to processing of personal data, defined as "any information relating to an identified or identifiable individual" [56]. The principles

Collection Limitation Principle. There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.

Data Quality Principle. Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date.

Purpose Specification Principle. The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.

Use Limitation Principle. Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with [the purpose specification principle] except (a) with the consent of the data subject; or (b) by the authority of law.

Security Safeguards Principle. Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorised access, destruction, use, modification or disclosure of data.

Openness Principle. There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

Individual Participation Principle. An individual should have the right (a) to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him; (b) to have communicated to him, data relating to him within a reasonable time; at a charge, if any, that is not excessive; in a reasonable manner; and in a form that is readily intelligible to him; (c) to be given reasons if a request made under subparagraphs (a) and (b) is denied, and to be able to challenge such denial; and (d) to challenge data relating to him and, if the challenge is successful to have the data erased, rectified, completed or amended.

Accountability Principle. A data controller should be accountable for complying with measures which give effect to the principles stated above.

Fig. 24.3 Basic principles of national application, Part 2 of the OECD privacy guidelines [56]

encourage parsimonious data collection and use, emphasize autonomy and consent, and promote data quality and security.

The Privacy Guidelines address data transfer on an international basis and—especially since the 2013 revision—the main provisions on transborder data flows include aspects of both the accountability and adequacy principles, also adding a principle of proportionality between risks and benefits (see Fig. 24.4).

Interestingly, unlike many of the subsequent international data privacy instruments, the Privacy Guidelines do not establish any “sensitive” categories of data subject to heightened privacy requirements.

16. A data controller remains accountable for personal data under its control without regard to the location of the data.
17. A Member country should refrain from restricting transborder flows of personal data between itself and another country where (a) the other country substantially observes these Guidelines or (b) sufficient safeguards exist, including effective enforcement mechanisms and appropriate measures put in place by the data controller, to ensure a continuing level of protection consistent with these Guidelines.
18. Any restrictions to transborder flows of personal data should be proportionate to the risks presented, taking into account the sensitivity of the data, and the purpose and context of the processing.

Fig. 24.4 Basic principles of international application: free flow and legitimate restrictions, Part 4 of the OECD privacy guidelines [56]

24.3.2 *The Council of Europe Convention 108*

In 1981, the year after the OECD Privacy Guidelines were released, the Council of Europe adopted its Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108) [18]. While the Council of Europe and the OECD coordinated their efforts in drafting their respective instruments, Convention 108 differs from the OECD framework not least in that it is the only legally binding international data protection instrument. Convention 108 [18, 20, 21] is technically a multilateral treaty dealing with data privacy, and is currently in force in all 47 member states of the Council of Europe except for San Marino and Turkey (the latter having signed but not yet ratified it). Furthermore, Article 23 of the Convention uniquely allows any state in the world to become a signatory, giving its privacy principles global potential (Council of Europe 1981). In 2013, the Convention entered into force in Uruguay, the first country outside of Europe to join; Morocco has received an invitation to join as well.

Since 2001, signatories have had the option of opting in to a set of added rules contained in an Additional Protocol. The rules in this protocol are now in force in more than two-thirds of the Council of Europe member countries and in Uruguay. A modernization process aimed at making significant amendments to Convention 108 was initiated in 2011 and remains ongoing.

The Convention requires that each signatory enact national laws to implement its data protection measures before the convention comes into force in the country. Although states are still allowed a measure of flexibility in their individual implementation, it is much narrower than that allowed by the OECD Privacy Guidelines.

Convention 108 expressly prohibits processing medical data unless national law provides “appropriate safeguards” (Article 6). Consequently, it is unlawful to process medical data about a person absent a legitimate basis for doing so, such as a doctor-patient relationship or the explicit consent of the data subject. A series of Council of Europe recommendations from the early 1990s on genetic testing

provide that genetic information should be processed in conformity with basic data privacy principles [10]. These recommendations are supplemented by the Council of Europe's 1997 Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine (otherwise known as the Oviedo Convention), which provides at Article 10(1) that "everyone has the right to respect for private life in relation to information about his or her health".

Moreover, the Council of Europe's Recommendation No. R (97) 5 on the Protection of Medical Data [19] applies the principles of Convention 108 to data processing in the medical field in more detail. Recommendation No. R (97) 5 explicitly acknowledges that scientific research may justify conserving personal data even after they have been used for the immediate purpose for which they were collected, although conservation usually requires their anonymisation. Article 12 of Recommendation No. R (97) 5 also sets out detailed proposed regulations to govern situations in which research using anonymised data would be impossible.

Convention 108's default position on data transfer—which it refers to as “data flows”—between countries in which the Convention is in force is that such transfer should not be restricted for privacy reasons [18]. Restrictions on personal data flows are allowed, however, either when the data pass through an intermediary that is not subject to the Convention, or insofar as a member country's legislation “includes specific regulations for certain categories of personal data . . . , because of the nature of those data . . . , except where the regulations of the other [member country] provide an equivalent protection” (Article 12(3)(a) [18]). Medical data are an obvious example of a “certain category” that national laws can nonetheless restrict due to their nature.

The Additional Protocol establishes that data flows to countries where the Convention is not in force are subject to the adequacy approach. “Adequacy” in this context is a functional concept that requires that the data protection regime in the importing country afford a sufficient level of protection, judged according to both the intended data processing activity (e.g., nature of the data, purpose and duration of the processing operation or operations), and the legal regime or measures applicable to the data recipient (e.g., general and sectoral rules of law, professional requirements and security measures). The Additional Protocol prohibits data transfer to non-signatories of the Convention unless either the receiving “State or organisation ensures an adequate level of protection for the intended data transfer,” or the data controller provides for safeguards that have been “found adequate by the competent authorities according to domestic law” (Article 2 [20]). The Additional Protocol also allows an exception for transfers that serve “legitimate prevailing interests, especially important public interests” or “specific interests of the data subject”, but only when domestic law provides for this (Article 2(2)(a) [20]). The modernization proposal, in its form at the time of writing this chapter, would incorporate a modified version of this aspect of the Additional Protocol into the Convention itself. Instead of requiring “adequate” protection, however, the proposed amendment demands “appropriate” protection, which it defines as follows:

An appropriate level of protection can be ensured by: (a) the law of that State or international organisation, including the applicable international treaties or agreements, or (b) ad hoc or approved standardised safeguards provided by legally binding and enforceable instruments adopted and implemented by the persons involved in the transfer and further processing (Council of Europe 2012:6 [21]).

Unless the measures described in subsection (b) above are involved, the proposed amendment would allow restrictions on transfer even between parties to the Convention if they originate from a country “regulated by binding harmonisation rules of protection shared by States belonging to a regional international organisation” [21]. The proposed amendment reformulates the public-interest and specific-interest exceptions mentioned above, and adds another exception to restrictions on transfer if “the data subject has given his/her specific, free and [explicit/unambiguous] consent, after being informed of risks arising in the absence of appropriate safeguards” [21].

24.3.3 The European Union Data Protection Directive 95/46

Arguably the most globally influential data privacy scheme is the European Union’s 1995 Directive on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (Directive 95/46 [28]).

Although EU members are already guaranteed to possess a degree of data privacy law harmonisation by virtue of all being signatories to Convention 108, Directive 95/46 seeks to elaborate on its principles and to promote harmonisation at the national level. Moreover, through its Charter of Fundamental Rights of the European Union, which came into force in 2009 with the Treaty of Lisbon, the EU explicitly conceives of personal data protection as a freestanding, fundamental right held by everyone (European Union 2010:Article 8(1) [29]). However, the right to respect for private life, as recognized in the 1950 European Convention on Human Rights and the Charter of Fundamental Rights of the European Union, and the right to data protection, as recognized in the Charter, are not absolute rights. As with all fundamental rights, they must be balanced with other rights, including academic freedom and freedom of scientific research, which is recognized in Article 13.

Unlike Convention 108, the privacy measures in Directive 95/46 are not directly enforceable, but the Directive is nonetheless legally binding in the 28 EU member states and the three European Economic Area (EEA) member countries. Directive 95/46 requires that each country establish data privacy laws implementing the Directive’s privacy measures, although EU Directives do allow member states to retain some discretion to implement measures in a way that accords with each national legal tradition. The Directive has accordingly been transposed into national law by all EU member states; by the three additional EEA member countries of Iceland, Liechtenstein, and Norway; and it has been taken up by Switzerland in a parallel law.

The Directive defines personal data as “any information relating to an identified or identifiable natural person ... who can be identified, directly or indirectly,

in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity” (European Union 1995:Article 2(a) [28]). It applies to all such data, irrespective of the citizenship or residency of the data subjects, with a few exemptions, including data used for purposes of national security.

Article 8(1) requires that states adopt a basic prohibition on processing categories of sensitive data, including “data concerning health” [28]. Among the exceptions to the prohibition are situations in which the consent of the data subject is given, situations of “substantial public interest”, and an exception in article 8(3), “where processing . . . is required for the purposes of preventive medicine, medical diagnosis, the provision of care or treatment or the management of health-care services, and where those data are processed by a health professional subject . . . to . . . professional secrecy or by another person also subject to an equivalent obligation of secrecy” [28]. Because biomedical research is absent from this list, unless a research project is justified by “substantial public interest” (which is undefined), it seems that the Directive precludes national laws from ever allowing research data to escape the basic prohibition. Nor is the effect of Article 8 on scientific research tempered by Article 13, which allows for some privacy obligations to be softened, including when “data are processed solely for the purposes of scientific research” (European Union 1995 [28]). The Directive’s only other reference to “scientific research”, a concept which it never defines, allows exemptions from notification and fair processing duties where these would be impossible or impractical when secondary use is made of data for “historical or scientific research” purposes (European Union 1995:Article 11(2) [28]).

As a result of the various instruments mentioned, laws in Europe that apply to biomedical research and processing of medical data are complex. Different areas of law, such as data privacy, tissue regulation, as well as biomedical research regulations can all impact how medical data must be handled. There are considerable differences in the way that European data privacy principles are implemented in national law. In particular, there are varying definitions of “personal data”, requirements for pseudonymisation, and rules on processing of data for biomedical research [48]. As Briceno Moraia and colleagues note, the Directive does not always contain all the relevant data privacy rules: “Most countries also have a specific legislation on medical research that must be read along with the data protection legislation to provide a comprehensive picture of the legal requirements that apply to medical research” [48].

The Directive defines consent as “any freely given specific and informed indication of . . . wishes by which the data subject signifies his agreement to personal data relating to him being processed” (European Union 1995:Article 2(h) [28]). But the practical scope of consent remains unclear. When, for example, during a long-term biomedical research study, is a fresh consent required? Even after the 2011 publication of a thirty-eight page opinion on such questions entitled “On the Definition of Consent” by the EU’s data protection working committee, the answer remains unclear [4].

Article 25 of the Directive governs data transfer to countries outside of the EU and EEA. Article 25(1) adopts the adequacy model, and Article 25(2) allows such

transfer only if adequate protection would be provided as assessed in the context of the particular transfer. The Directive provides a mechanism that allows transfers to be deemed adequate where the European Commission has previously determined that legislation governing the transferee offers adequate protection. To date, the European Commission has found that data privacy laws in thirteen jurisdictions provide adequate protection.

This said, Article 26 provides several exceptions that allow data transfer in other situations. Article 26(1) provides for exceptions either with the “explicit consent” of the data subject, to comply with certain legal obligations, or where “necessary in order to protect the vital interests of the data subject” [28]. Article 26(2) allows transfer absent a European Commission adequacy decision where the data “controller adduces adequate safeguards” [28]. The principal safeguards that data controllers may rely on are contractual data protection guarantees, binding corporate rules (BCRs), and the EU-US Safe Harbor Agreement, the latter constituting an agreement that allows US companies to self-certify compliance with specific privacy principles and subject themselves to the enforcement supervision of the US Federal Trade Commission in order to receive personal data from Europeans. It is worth noting that publication of personal data is not considered to be transborder data flow; only communication which is directed at specific recipients is captured by the concept. Thus, data published in an online public register is not transborder data flow per se.

The Directive is now set to be superseded by a new General Data Protection Regulation, the first draft of which was released by the European Commission in January 2012 [25, 27]. Unlike a Directive, a Regulation in principle allows no room for legal manoeuvring by member states. The law will automatically be transposed and applicable across the EU and EEA member states. The Regulation’s legislative process involves a complicated interaction between the EU Parliament, Council, and Commission—and ultimately their agreement. The EU Parliament passed its proposed text of the Regulation in March 2014 with over 95 % support, but an ongoing succession of controversial amendments, with varying approaches to research and health data, make it unclear what exactly the final result will look like when the completed Regulation finally emerges from the trilateral negotiations between the three EU bodies.

The Regulation will likely extend the scope of EU data protection to apply to any company, worldwide, which processes the personal data of EU residents. Existing Data Protection Authorities (DPAs), which oversee data controllers in certain circumstances, will be amalgamated into public “supervisory authorities” to be established in each member state (Article 46 in drafts to date).

Although data transfer rules would remain functionally similar to what exists already, the adequacy approach would be supplemented in the law by three alternative means to allow data transfer outside the EU. First, data controllers would still be able to rely on an adequacy decision made by the European Commission, including those made pursuant to Directive 95/46 (Article 41 in drafts to date). Second, they would still be entitled to guarantee appropriate safeguards, which must now also apply to any subsequent transfers. Examples of appropriate safeguards

considered by the Regulation are BCRs; standard data protection clauses adopted by the Commission and potentially also by a supervisory authority; an approved code of conduct or certification mechanism, either one of which must include binding obligations in the third country; or authorisation by the relevant supervisory authority of contractual guarantees or analogous administrative arrangements between public bodies. Third, data transfer would also be allowed by the application of a listed exception, referred to as a derogation (Article 44 in drafts to date).

The Regulation would also place a *prima facie* blanket restriction on processing special categories of data—including “genetic data or data concerning health” (Article 9 in drafts to date). In the Commission’s draft Regulation from 2012, one set of exceptions to this general prohibition was established for health data (Article 81 in drafts to date) and another for scientific purposes (Article 83 in drafts to date). These exceptions have been some of the most controversial and frequently amended provisions in the draft Regulation. The drafts appear to intend that the exceptions relevant for biomedical research are those specified for research purposes rather than those established for health data.

In June 2015 [26], the Justice and Home Affairs Council of the European Council agreed on its General Approach on the content of the entire Regulation. The European Council’s version amended the Commission’s initial draft from 2012. It removed Article 81 (processing of personal data for health-related purposes), suggesting the Article was superfluous and noting that their proposed version of Article 9 enshrines the basic idea, previously expressed in Article 81, that sensitive data such as medical data may be processed for, among other purposes, medicine (including processing of genetic data necessary for medical purposes), healthcare, public health and other public interests, subject to certain appropriate safeguards based on European Union law or the national law of a member state, as well as for scientific (e.g., research) purposes, subject to the conditions and safeguards referred set forth in Article 83. At the same time, the Council’s text provides a consolidated Article 83 that addresses all types of derogations (i.e., archiving, scientific, statistical and historical purposes) from the general prohibition on processing special categories of personal data. The draft texts retain Article 6(2) from the Commission’s initial 2012 proposal, which clarifies that processing of (non-sensitive) personal data that is necessary for scientific purposes (including research) is itself a lawful purpose, subject to the conditions and safeguards of Article 83.

The Council’s text remarks in Article 83 that where personal data are processed for archiving, scientific, statistical or historical purposes, appropriate safeguards for the rights and freedoms of the data subject must be in place either through EU or Member State law. The appropriate safeguards should be such as “to ensure that technological and/or organisational protection measures pursuant to this Regulation are applied to the personal data, to minimise the processing of personal data in pursuance of the proportionality and necessity principles, such as pseudonymising the data, unless those measures prevent achieving the purpose of the processing and such purpose cannot be otherwise fulfilled within reasonable means” (European Council 2015:195 [26]). This explicit reference to and endorsement of pseudonymi-

sation is a similar approach to the Commission's initial draft, but widens the scope of potential safeguards beyond pseudonymisation.

This widening is essential, as pseudonymisation provides little protection on its own. The technique is only effective alongside safeguards that ensure that pseudonymised data will not be accessed by a person who will attempt to re-identify them. However, this approach does a certain amount of injustice to its own proportionality principle, which is abandoned when processing purposes "cannot be otherwise fulfilled within reasonable means". This approach would allow the most sensitive data to be processed using the most unreasonable means to achieve the most trifling benefit, so long as no safer approach exists. A more natural strategy would retain a proportionality criterion in these circumstances, allowing processing to occur only when the anticipated benefits of processing clearly outweigh the risks.

Analysis of the Regulation is necessarily incomplete and speculative as further negotiation in EU legislative and executive bodies continues. It remains to be seen how exactly the final Regulation will impact the processing of medical data for research purposes, clinical purposes, or otherwise, especially as it relates to issues of anonymisation, pseudonymisation, and consent [36].

24.3.4 UK Data Protection Act 1998

The Data Protection Act 1998 (DPA) [65] is the UK's implementation of EU Directive 95/46. The United Kingdom has no specific statute for patient rights and medical research [48]. The DPA applies to personal data, meaning "data which relate to a living individual who can be identified (a) from those data, or (b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller" [65]. It also defines a subcategory of "sensitive personal data", meaning "personal data consisting of information as to ... [the person's] physical or mental health or condition" [65].

The DPA confers on data subjects certain rights of access, rights to prevent processing likely to cause damage or distress or for direct marketing purposes, and rights to be free from serious effects of decisions made automatically by personal data processing. It requires data controllers to register with an Information Commissioner whose position the Act creates. At the core of the DPA are eight privacy principles (see Fig. 24.5).

The situations that allow processing of personal data for the purposes of DPA principle 1(a) are where the data subject has given consent or where the processing is "necessary" either "to protect the vital interests of the subject," when certain legal obligations—including some contractual obligations—are at stake, for the exercise of the office of certain public officials, or "for the purposes of legitimate interests pursued by the data controller" or the transferee (United Kingdom 1998:Schedule 2 [65]).

The conditions that are additionally required by DPA principle 1(b) where sensitive data—such as medical data—are involved are similar, though often narrower. They include cases where the data subject has given "explicit" consent; a

1. Personal data shall be processed fairly and lawfully and, in particular, shall not be processed unless – (a) at least one of the conditions in Schedule 2 is met, and (b) in the case of sensitive personal data, at least one of the conditions in Schedule 3 is also met.
2. Personal data shall be obtained only for one or more specified and lawful purposes, and shall not be further processed in any manner incompatible with that purpose or those purposes.
3. Personal data shall be adequate, relevant and not excessive in relation to the purpose or purposes for which they are processed.
4. Personal data shall be accurate and, where necessary, kept up to date.
5. Personal data processed for any purpose or purposes shall not be kept for longer than is necessary for that purpose or those purposes.
6. Personal data shall be processed in accordance with the rights of data subjects under this Act.
7. Appropriate technical and organisational measures shall be taken against unauthorised or unlawful processing of personal data and against accidental loss or destruction of, or damage to, personal data.
8. Personal data shall not be transferred to a country or territory outside the European Economic Area unless that country or territory ensures an adequate level of protection for the rights and freedoms of data subjects in relation to the processing of personal data.

Fig. 24.5 The UK data protection principles, Part 1 of Schedule 1 of the Data Protection Act 1998 [65]

narrowed “vital interests” exception that nonetheless extends to interests of another person; a much narrower “legitimate interests” exception; and where processing is done for “medical purposes” by a health professional or a person with a similar duty of confidentiality (United Kingdom 1998:Schedule 3 [65]). Medical purposes here “includes the purposes of preventative medicine, medical diagnosis, medical research, the provision of care and treatment and the management of healthcare services.”

This list includes medical research, which is conspicuously absent from the listed exceptions permitted by Article 8(3) of EU Directive 95/46 [28]. A report commissioned by the European Commission in 2010 expressed the view that this mismatch renders the DPA “blatantly in violation of the Directive” [39].

The DPA permits the use of personally identifiable or pseudonymised data for research, as it contains a limited but important exemption for research purposes (section 33). The exemption allows for personal data to be used, provided that (a) the data are not processed to support measures or decisions with respect to particular individuals; (b) the data are not processed in such a way that substantial damage or substantial distress is, or is likely to be, caused to any individual; and (c) neither the results of the research nor any resulting statistics are made available in a form which identifies any individuals. If identification of individuals does take place, research participants must not have previously been assured that only anonymised data would be published.

It should be noted that the Data Protection (Processing of Sensitive Personal Data) Order 2000 [66] also governs the circumstances in which sensitive data may be processed. Paragraph 9 states that processing of sensitive data may occur where the processing (a) is in the substantial public interest; (b) is necessary for research purposes (which has the same meaning as section 33 of the DPA); (c) does not support measures or decisions with respect to any particular data subject otherwise than with the explicit consent of that data subject; and (d) does not cause, nor is likely to cause, substantial damage or substantial distress to the data subject or any other person (United Kingdom 2000:Paragraph 9 [66]).

24.3.5 Canadian Privacy Legislation

In contrast to European jurisdictions, Canada and the United States, two countries with federal systems of government, take a more fragmented approach to data privacy. Two distinct federal Canadian laws share the primary responsibility for privacy protection. The 1983 Privacy Act [12] governs the handling of personal information by the federal government, and the 2000 Personal Information Protection and Electronic Documents Act (PIPEDA) [13] applies to private sector commercial organisations.

To complicate matters, all ten Canadian provinces and its three territories each have their own data privacy laws. Nearly all also have specific health-information privacy laws, and all have health-information privacy provisions tailored specifically to the health sector. Provincial private-sector data privacy law prevails over PIPEDA where an order declaring that the former provides “substantially similar” protection has been made by the federal government (Canada 2000:s.26(2)(b) [13]). Currently, private-sector data privacy legislation in Quebec, British Columbia, and Alberta prevails over PIPEDA throughout those provinces, and health-information privacy legislation in the three provinces of Ontario, New Brunswick, and Newfoundland and Labrador prevails over PIPEDA as it applies to health data controllers.

PIPEDA is inspired by the OECD Guidelines and is seen as a compromise between the fundamental rights-based European approach and the United States tendency toward industry self-regulation and multiple sector-specific privacy instruments. The European Commission determined in 2002 that PIPEDA provides adequate protection for the purposes of Directive 95/46 [28] such that personal data of Europeans can flow to commercial organisations in Canada.

PIPEDA defines personal information as “information about an identifiable individual, but does not include the name, title or business address or telephone number of an employee of an organisation” (Canada 2000:s.2(1) [13]). Its ten privacy principles derive largely from those of the OECD Guidelines (Canada 2000:Schedule 1 [13]). It follows the accountability and organisational approach to data transfer, and holds the data controller responsible for the personal data in its possession or custody, including data that has been transferred to a third party for processing. Data controllers must use contractual or other means (e.g., privacy

policies, training programs, mechanisms to enforce compliance) in order to provide a comparable level of protection while the data are being processed by a third party. PIPEDA does not require prior approval for data transfers and only provides accountability after the fact.

Provincial laws, and especially provincial health information privacy laws, provide a robust framework that is tailored with a view to appropriately managing the disclosures allowable in health research. Data controllers may disclose medical data without an individual's consent to researchers provided that researchers submit a research plan to the data controller and have in place (and make transparent) privacy practices, policies, and procedures approved by a duly authorised body (e.g., data protection authority) and that the data disclosed to researchers is either in de-identified form or in identifiable form with approval of a research ethics committee. These committees are multidisciplinary panels of medical, legal, and ethics experts. They must generally evaluate and approve proposed research projects, sometimes imposing additional conditions on the research plan, such as safeguards for medical data. The approval process usually requires that the panel assess relevant considerations such as the public interest in the research relative to the privacy risks in the circumstances; the practicability of obtaining consent; the adequacy of privacy safeguards adopted; and whether the research objectives can be reasonably met without using personal medical data.

24.3.6 *The HIPAA Privacy Rule*

As noted above, the US takes a sectoral approach to privacy legislation. Even the most general US privacy laws, such as the Privacy Act of 1974 and the Computer Matching and Privacy Act, apply only to federal agencies, and so have limited impact on medical data and their use by clinicians and researchers. Most US privacy law is extremely narrow in scope, such as the 1988 Video Privacy Protection Act, which prohibits video tape service providers from disclosing personally identifiable information (e.g., video tape rentals and sales records) concerning any consumer of such provider. Privacy provisions also often appear as incidental parts within a broader statute whose main purpose is unrelated to privacy. A privacy provision can be found, for example, within the chapter of the federal US Code [70, 71] that authorises the creation of the Public Health Service. A section in the chapter addressing "General provisions respecting effectiveness, efficiency, and quality of health services" contains within it a subsection on the protection of personal information obtained for research purposes by the National Centers for Health Services and for Health Statistics [title 42, section 242m(d)].

The most relevant US data privacy law for the purposes of this chapter is the 1996 Health Insurance Portability and Accountability Act (HIPAA) [70, 72], which specifically regulates health privacy. It was the first comprehensive US federal Department of Health and Human Services guideline for the protection of the privacy of "protected health information" (PHI). HIPAA's Privacy Rule, which was

adopted in 2002 and came into force in 2003, applies to “covered entities”, which include healthcare providers, health plans, and healthcare clearinghouses. The 2009 Health Information Technology for Economic and Clinical Health Act (HITECH Act) expanded HIPAA’s scope to include the “business associates” of these covered entities and their subcontractors, such as cloud service providers processing medical data. Additionally, in January 2013, the US Federal Register published omnibus amendments by the Department of Health and Human Services to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules. These modifications also include the final versions of the HIPAA regulation amendments mandated by the HITECH Act.

HIPAA provides that a covered entity may not use or disclose PHI except either (1) as permitted by the Privacy Rule, or (2) as authorised in writing by the individual who is the subject of the information (or the individual’s personal representative). HIPAA defines information as identifiable when “there is a reasonable basis to believe the information can be used to identify [an] individual” (United States 2014:§160.103 [70]). But HIPAA’s scope is circumscribed further, as it protects only individually identifiable health information, though this includes “demographic information collected from an individual” (United States 2014:§160.103 [70]). Health information, in turn, is information that “[r]elates to the past, present, or future physical or mental health or condition of an individual; the provision of healthcare to an individual; or the past, present, or future payment for the provision of healthcare to an individual” (United States 2014:§160.103 [70]).

Contrary to the claims of some commentators, the HIPAA Privacy Rule does bear on biomedical research. While it does not regulate biomedical researchers per se, when obtaining PHI (as defined in the Privacy Rule) from a covered entity to use in health research, researchers must nonetheless follow the provisions of the HIPAA Privacy Rule. Medical records-based research, in which the PHI come from documents or databases and not directly from participants, is also subject to the HIPAA Privacy Rule. A covered entity is permitted to use and disclose PHI for research purposes under several conditions (see Fig. 24.6).

HIPAA is one of very few data privacy laws in the world that address data de-identification in technical detail. It provides two options that allow data to be considered de-identified for its purposes. First, a data controller can obtain a detailed written opinion from a statistician assuring “that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information” (United States 2014:§164.514(b)(i) [71]). Because various components of this provision are not clearly defined—for instance, the necessary qualifications of the statistician—*data controllers have rarely risked relying on it*.

The second option, sometimes referred to as the “Safe Harbor” (not to be confused with the formal agreement referred to above that allows US companies to conform to the EU Directive 95/46’s extra-territorial data transfer provisions [72]), has been much more popular. The HIPAA Safe Harbor requires that the data controller meet three conditions. They must first remove seventeen specified fields from each record in the dataset (sometimes referred to as “The List”, see

- If research participants provide a written authorisation.
- If a privacy officer/board has granted a waiver of authorisation requirement.
- If the PHI have been de-identified.
- If the researcher uses a “limited data set” and a “data use agreement”.
- If legal permission to disclose the PHI is ongoing, or originated before HIPAA came into effect (e.g., in an informed consent form or an IRB waiver of informed consent).
- If it has been grandfathered by the HIPAA transition provisions for research on a descendant’s information if the researcher provides the required documentation.

Fig. 24.6 Conditions under which a covered entity is permitted to use and disclose PHI for research purposes [70]

1. Names;
2. All geographic subdivisions smaller than a State . . . except for the initial three digits of a zip code if . . . (1) The geographic unit formed . . . contains more than 20,000 people; and (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
3. All elements of dates (except year) . . . directly related to an individual . . . ; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
4. Telephone numbers;
5. Fax numbers;
6. Electronic mail addresses;
7. Social security numbers;
8. Medical record numbers;
9. Health plan beneficiary numbers;
10. Account numbers;
11. Certificate/license numbers;
12. Vehicle identifiers and serial numbers, including license plate numbers;
13. Device identifiers and serial numbers;
14. Web Universal Resource Locators (URLs);
15. Internet Protocol (IP) address numbers;
16. Biometric identifiers, including finger and voice prints; [and]
17. Full face photographic images and any comparable images[.]

Fig. 24.7 The seventeen HIPAA Privacy Rule de-identification fields (U.S. 2014: §164.514(b)(2)(i) [71])

Fig. 24.7). They must then also remove “any other unique identifying number, characteristic, or code” except for an optional re-identification keycode which, if present, must be created and used according to specific rules (United States 2014:§164.514(b)(2)(i)(R) [71]). Finally, the data controller must not in fact know

Table 24.1 HIPAA Safe Harbor example de-identification of a simple medical dataset prior to de-identification and following de-identification (changed/removed data in bold)

Social security no.	Occupation	Zip code	Age	Patient number	Diastolic blood pressure recorded on 25 July 2014
248-09-1593	Civil servant	24860	90	592-0969	72.5
418-12-0635	Restaurant cook	06351	95	593-9319	79.2
721-07-4426	Mayor	79937	85	590-0393	90.3

Re-identification no.	Occupation	Zip code	Age	Diastolic blood pressure recorded in 2014
1	Civil servant	24800	90	72.5
2	Restaurant cook	00000	90	79.2
3	REMOVED	79900	85	90.3

that the remaining “information could be used alone or in combination with other information to identify an individual who is a data subject” (United States 2014:§164.514(b)(2)(ii) [71]). Consider the de-identification example provided in Table 24.1: although a person’s occupation is neither in “The List” nor will it generally enable individual identification on its own, the third data subject’s occupation has been redacted because when the occupation of “mayor” is combined with the subject’s zip code, which places her in El Paso, Texas, the record offers (considerably) more than a reasonable chance of allowing the individual to be identified.

The HIPAA Safe Harbor de-identification framework has been criticized, principally on the grounds that without individually analysing the remaining fields, which the data controller is under no obligation to do, it will generally be impossible to know the likelihood of re-identification, which may be trivial. A report by the US Institute of Medicine found HIPAA was simultaneously too strict and not strict enough in protecting privacy:

The HIPAA Privacy Rule does not protect privacy as well as it should, and . . . impedes important health research. The . . . Privacy Rule (1) is not uniformly applicable to all health research, (2) overstates the ability of informed consent to protect privacy rather than incorporating comprehensive privacy protections, (3) conflicts with other federal regulations governing health research, (4) is interpreted differently across institutions, and (5) creates barriers to research and leads to biased research samples, which generate invalid conclusions. In addition, security breaches are a growing problem for health care databases (Institute of Medicine 2009 [40]).

This said, there is no consensus view on this point. In 2011, for example, a systematic literature review was conducted to identify and assess published accounts of re-identification attacks on de-identified health datasets. The literature review identified 14 accounts of re-identification attacks on ostensibly de-identified health information, but found that only two of the 14 attacks were made on datasets that

A limited dataset is protected health information that excludes the following direct identifiers of the individual or of relatives, employers, or household members of the individual:

1. Names;
2. Postal address information, other than town or city, State, and zip code;
3. Telephone numbers;
4. Fax numbers;
5. Electronic mail addresses;
6. Social security numbers;
7. Medical record numbers;
8. Health plan beneficiary numbers;
9. Account numbers;
10. Certificate/license numbers;
11. Vehicle identifiers and serial numbers, including license plate numbers;
12. Device identifiers and serial numbers;
13. Web Universal Resource Locators (URLs);
14. Internet Protocol (IP) address numbers;
15. Biometric identifiers, including finger and voice prints; and
16. Full face photographic images and any comparable images.

Fig. 24.8 HIPAA Privacy Rule limited dataset (U.S. 2014:§164.514(e)(2) [71])

had been properly de-identified in accordance with the HIPAA Safe Harbor requirements. The remaining 12 attacks were made on datasets that had failed to meet this standard of de-identification [24]. Such an anecdotal review cannot establish the appropriateness of a data protection measure, but the debate is ongoing as to whether the HIPAA Safe Harbor requirements appropriately address the concerns of data privacy and of biomedical researchers. With respect to genomic data, many commentators remark that because genomic data are inherently individuating, any de-identification processes will provide at best limited privacy protection. Indeed, the data in the NIH database of Genotypes and Phenotypes (dbGaP) are de-identified in accordance with both the HHS regulations for protection of human subjects and HIPAA Privacy Rule standards, and the NIH has additionally obtained a Certificate of Confidentiality for dbGaP as an added precaution because of the relative ease with which genomic data can be re-identified.

The HIPAA Privacy Rule allows a less strict variation on its Safe Harbor called a *limited dataset* (see Fig. 24.8) “only for the purposes of research, public health, or health care operations” that may be transferred by a data controller who “obtains satisfactory assurance . . . that the limited data set recipient will only use or disclose the protected health information for limited purposes” and undertakes to “not identify the information or contact the individuals,” among other requirements (United States 2014:§164.514(e) [71]). A limited dataset is similar to the de-identified Safe Harbor dataset but requires fewer identifiers to be removed. Limited datasets may include city, state, zip code, elements of a date, and other numbers, characteristics, or codes not listed as direct identifiers. Typically, limited datasets are utilised in multicentre studies when using fully de-identified data would not be

useful. They allow researchers and others to have access to dates of admission and discharge, birth and death, and five-digit zip codes or other geographic subdivisions other than street address. Limited datasets do not include specified direct identifiers of the individual's relatives, employers, or household members.

To use a limited dataset, a researcher must sign a “data use agreement” that limits who can use or receive the limited dataset. It requires that the researcher neither re-identify the data nor contact the research participant and that the researcher obtains satisfactory assurance that safeguards defined in the HIPAA rule will be used to prevent improper use or disclosure of the limited dataset.

24.4 Data Sharing Policies

Since 1992, a parallel current to data privacy legal regimes has emerged from the life sciences, particularly in genomics. The 1992 Guidelines for Access to Mapping and Sequencing Data and Material Resources, adopted by the US-based Joint Subcommittee on the Human Genome, recommended no more than a six-month delay between the time genomic data are generated and made public. Rapid sharing was recognized as particularly important within the field of genomics, but a balance was struck to allow researchers “time to verify the accuracy of their data and to gain some scientific advantage from the effort they have invested” [53].

Revised statements on rapid release of genomic and later other health-related data were adopted in 1996, 2003, 2008, 2009, and most recently in 2011 in the Joint Statement by Funders of Health Research. The rapid data release principle has been expanded in some of these statements to apply to subfields such as proteomics and metabolomics. Twenty-four hours has, in some cases, come to be the accepted delay for prepublication release of generated data, though longer delay (e.g., several weeks or months) for quality control and other purposes is also generally accepted.

Despite the increasing nuance and expansion to non-genomic data fields, the rapid data-release standards in many data sharing policies have been maintained, partly because large sets of genomic or “-omic” data have an inherently public character (particularly if publicly funded) which demands use for public benefit, and partly because other mechanisms have been developed to safeguard the initial researchers’ (i.e., data producers’) interests, often by reserving the opportunity to publish first on the data, even if other researchers have already begun to make research use of and findings on the data. As the brief following discussion of US, Canadian, and UK data sharing policies shows, the principle of rapid data release and open data sharing has now often become a core element of successful biomedical research funding, but the ways in which it is operationalized in these policies can differ subtly.

As with data privacy regulations, data sharing policies of major biomedical research funding organisations have aimed to promote both open data flows and appropriate privacy protections. But data sharing policies depart from data privacy regulations in that rather than aiming to strike an appropriate balance between

privacy and data sharing—and with the force of law—they instead aim to minimise or eliminate impediments to rapid data sharing that are not directly related to privacy concerns. For example, these policies aim to counteract the inclination of researchers to not share their data because they are overly protective of their work, because they believe too many resources are required to make the data shareable, because they believe their data would not be useful to others, or because they simply cannot be bothered. As discussed below, this can result in placing medical data controllers or researchers in a double bind [55]: they are exhorted to generate, use and extend access to data (because doing so is expected to advance research and healthcare); and, at the same time, they are required by law to protect privacy.

24.4.1 US National Institutes of Health

The NIH is the primary institution in the US responsible for biomedical research and funding. It has many institutional data sharing policies. Among the most significant for genomic research is the NIH's 2007 Policy for Genome-Wide Association Studies (GWAS Policy) [49]. The GWAS Policy "strongly encourages the submission of curated and coded phenotype, exposure, genotype, and pedigree data, as appropriate, to the NIH GWAS data repository as soon as quality control procedures have been completed" to be made available to other researchers [49].

As a trade-off, the GWAS Policy provides that "investigators who contribute data to the NIH GWAS data repository will retain the exclusive right to publish analyses of the dataset for a defined period of time" up to a "maximum period of . . . twelve months from the date that the GWAS dataset is made available for access through the NIH GWAS data repository" [49]. "During this period of exclusivity, the NIH will grant access through the DACs to other investigators, who may analyze the data, but are expected not to submit their analyses or conclusions for publication during the exclusivity period" [49].

The GWAS Policy requires that "in order to minimise risks to study participants, data submitted to the NIH GWAS data repository will be de-identified and coded" according to the HIPAA Safe Harbor Privacy Rule [49]. After a paper was published showing that the GWAS data were re-identifiable (Homer et al. [38]), the NIH began to require that researchers seek approval prior to gaining access to individual-level dbGaP data [50].

In August 2014, the NIH released its much anticipated Genomic Data Sharing Policy [3, 51]. This new policy, which replaces the GWAS Policy, still requires extensive sharing of data, but, in recognition of the emergence of other large, well-established public databases emerging around the globe, "NIH-designated data repositories need [no longer] be the exclusive source for facilitating the sharing of genomic data, that is, investigators may also elect to submit data to a non-NIH-designated data repository in addition to an NIH-designated data repository. However, investigators should ensure that appropriate data security measures are in place, and that confidentiality, privacy, and data use measures are consistent with

the” Genomic Data Sharing Policy [51]. A supplement to the Genomic Data Sharing Policy separates data into five levels according to the degree of processing that has been carried out on them, and indicates NIH expectations for data submission and data release timelines for each level [52] (see Table 24.2).

A novel element of the Genomic Data Sharing Policy is its focus on what may be termed “specifically broad consent” (personal communication, Bartha Maria Knoppers), that is, an explicit expectation from the funder that “investigators generating genomic data . . . seek consent from participants for future research uses and the broadest possible sharing” [51]. Clearly, a necessary counterpart to allowing investigators to actively and explicitly seek participants’ consent for the broad sharing of their research data is that the data must be subject to strong privacy protections, and perhaps also recognition and communication to participants that privacy risks persist despite the de-identification of their genomic data.

24.4.2 Canadian Data Sharing Policies

Two important Canadian data sharing policies can be compared to the recent NIH Genomic Data Sharing Policy. The Canadian Institutes of Health Research (CIHR) is the major public federal agency that funds Canadian health research. In 2013, it released its CIHR Open Access Policy, which affected all projects it funded, not just those related to medical data [14]. In 2015, this policy was replaced by the Tri-Agency Open Access Policy on Publications, which applies to CIHR, the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC). The policy aims to improve access to the results of research funded by these agencies and to increase the dissemination and exchange of research results.

Under the Tri-Agency Open Access Policy, all CIHR grant recipients are required to “deposit bioinformatics, atomic, and molecular coordinate data into the appropriate public database, as already required by most journals, immediately upon publication of research results” [32]. This policy notably lacks a data sharing plan or publication embargo period. Instead, it requires only that limited types of medical data are deposited into a public database “immediately” upon publication of research results. While embargo restrictions may be difficult to police, they nonetheless serve to balance improved “lead-time by restricting the ability of users to publish conclusions based on the data” with making large quantities of data rapidly available to the scientific community to foster scientific advancement [16]. Further, only a subset of the total research data generated by CIHR grant recipients must necessarily be publicly deposited.

Genome Canada, the major federal Canadian funder of genomic research, published its Data Release and Resource Sharing policy in 2008. In line with the NIH data sharing policies, it declares that “Genome Canada-funded projects must . . . share data and resources in a timely fashion with minimal or no restrictions”

Table 24.2 Deadlines for data submission and release in a supplement to the NIH Genomic Data Sharing Policy, which apply to all large-scale, NIH-funded genomics research

Level processing	General description of data	Example data types	Data submission expectation	Data release timeline
0	Raw data generated directly from the instrument platform	Instrument image data	Not expected	N/A
1	Initial sequence reads, the most fundamental form of the data after the basic translation of raw input	DNA sequencing reads, ChIP-Seq reads, RNA-Seq reads, SNP arrays, Array CGH	Not expected, except not later than the time of initial publication for non-human, de novo sequence data (unless included with Level 2 aligned sequence files)	<ul style="list-style-type: none"> 1. N/A, for human data 2. Generally no later than the time of initial publication, for non-human de novo sequence data
2	Data after an initial round of analysis or computation to clean the data and assess basic quality measures	DNA sequence alignments to a reference sequence or de novo assembly, RNA expression profiling	<ul style="list-style-type: none"> 1. Project specific, for human data, but generally within 3 months after data generation 2. Generally no later than the time of initial publication, for non-human data 	<ul style="list-style-type: none"> 1. Up to 6 months after data submission or at the time of acceptance of initial publication, whichever occurs first, for human data 2. Generally no later than the time of initial publication, for non-human data
3	Analysis to identify genetic variants, gene expression patterns, or other features of the dataset	SNP or structural variant calls, expression peaks, epigenomic features	<ul style="list-style-type: none"> 1. Project specific, generally within 3 months after data generation, for human data 2. Generally no later than the time of initial publication, for non-human data 	<ul style="list-style-type: none"> 1. Up to 6 months after data submission or at the time of acceptance of the first publication, whichever occurs first, for human data 2. Generally no later than the time of initial publication, for non-human data

Table 24.2 (continued)

Level	General description of data processing	Example data types	Data submission expectation	Data release timeline
4	Final analysis that relates the genomic data to phenotype or other biological states	Genotype-phenotype relationships, relationships of RNA expression or epigenomic patterns to biological state	1. Data submitted as analyses are completed, for human data 2. No later than the time of initial publication, for non-human data	1. Data released with publication, for human data 2. No later than the time of initial publication, for non-human data

The deadlines vary according to five different levels of data processing [52]

Genome Canada expects researchers to share data and resources as rapidly as possible. Where the goal of the project is to produce data or resources for the wider scientific community the project must follow the data release and resource sharing principles of a “Community Resource Project”, defined as “a research project specifically devised and implemented to create a set of data, reagents or other material whose primary utility will be as a resource for the broad scientific community.” This definition and the associated data release and resource sharing principles were developed at a meeting held in January 2003 in Fort Lauderdale.

Genome Canada encourages the application of the principles of rapid, pre-publication data release to other types of projects and is working with other research funders to promote this practice.

Genome Canada recognizes publication as a vehicle for data release, and, at a minimum, expects data to be released and shared no later than the original publication date of the main findings from any datasets generated by that project. For large datasets that are collected over several discrete time periods or phases, it is reasonable to expect that the data be released in phases as they become available or as main findings from a research phase are published. However, at the conclusion of a project, all data must be released without restriction.

Fig. 24.9 Extract from the genome data release and resource sharing policy [31]

[31]. The policy requires that all funding applicants submit a detailed data- and resource-sharing plan as part of their application, which must conform to the policy (see Fig. 24.9).

Like the Tri-Agency Open Access Policy, there is no direct discussion of an embargo period, and like the NIH data sharing policies, Genome Canada applicants are not required to make their data and resource-sharing plan public. The policy provides relatively scant guidance for sharing data in a privacy-promoting manner. Researchers need only address the question “If the data could be of a potentially sensitive nature, how will this be handled?” which, absent any connection to relevant data privacy regulation and ethical guidelines, suggests researchers are free to disregard, at least for Genome Canada’s purposes, the equally important issue of how they will handle all personal medical data in their control in a way that protects and promotes privacy-related interests.

24.4.3 Wellcome Trust (UK)

The Wellcome Trust is the UK’s largest non-governmental funder of scientific research. In 2010 it released its Policy on Data Management and Sharing [76], which is reproduced in its entirety in Fig. 24.10.

The Wellcome Trust clearly exhorts its funding recipients to share research data as widely and as quickly as possible. Like the NIH and Genome Canada data sharing policies, it requires applicants to submit a data sharing plan. A plan is necessary only

1. The Wellcome Trust expects all of its funded researchers to maximise the availability of research data with as few restrictions as possible.
2. All those seeking Wellcome Trust funding should consider their approach for managing and sharing data at the research proposal stage. In cases where the proposed research is likely to generate data outputs that will hold significant value as a resource for the wider research community, applicants will be required to submit a data management and sharing plan to the Wellcome Trust prior to an award being made.
3. The Wellcome Trust will:
 - review data management and sharing plans, and any costs involved in delivering them, as an integral part of the funding decision.
 - work with grant holders on an ongoing basis to support them in maximising the long-term value of key datasets resulting from their research.
4. The Wellcome Trust expects all users of research data to acknowledge the sources of their data and to abide by the terms and conditions under which they accessed the original data.
5. The Wellcome Trust will foster an environment that enables researchers to maximise the value of research data. Specifically, we will work in partnership with others to:
 - ensure that key data resources are developed and maintained for use by the research community.
 - recognise the contributions of researchers who generate, preserve and share key research datasets.
 - develop best practice for data sharing in different fields – recognising that different data types raise distinct issues and challenges.

Fig. 24.10 Wellcome Trust policy on data management and sharing [76]

if the research “is likely to generate data outputs that will hold significant value as a resource for the wider research community”, a vague condition, but one that seems to apply to certain types of medical data.

Two points are notable about the Wellcome Trust data sharing policy: first, its lack of any statement about data privacy, and second, the explicit commitment made by the Trust (1) to provide resources to researchers to help sustain key datasets, and (2) to foster a “share and share alike” culture encouraging sharing and recognition of those who have made their data available for sharing. Thus, while the Wellcome Trust’s policy goes further than the NIH and Genome Canada policies in encouraging (and financially supporting) both widespread sharing of data and appropriate attribution for data producers, the latter being an especially important incentive for many in the biomedical research community, it does so at the expense of explicitly addressing the issue of how to share research data in ways that maximise public benefit (i.e., generate improvements in health) and minimise public harm (i.e., violations of privacy interests).

Taken together, the review of data sharing policies in this section illustrates that several key biomedical research funders explicitly support a data sharing culture. At the same time, the policies do not always address the need for research

investigators to ensure that appropriate data security measures are in place, nor that confidentiality, privacy, and data-use measures are consistently addressed across data sharing policies (although the recent NIH Genomic Data Sharing Policy may inaugurate future improvements in this respect). Exhorting researchers to share their data while failing to discuss data privacy issues fosters tension between widespread data sharing and protecting medical data privacy.

Some data sharing policies explicitly curtail any of their data sharing requirements that would infringe data privacy laws, and understandably so, as such policies have no authority to override legal rules (except to the extent provided for by the law itself, such as when a data subject explicitly consents). For this reason, policy developments in this area depend crucially on the content of data privacy law and thus, conversely, legislators and regulators ought to consider the benefits of responsible medical data sharing when drafting and administering data privacy laws and amendments that otherwise might unduly restrict the ability to share medical data within and across jurisdictions. In an age in which medical data are shared across borders and in places of divergent data privacy regulation, it remains to be determined whether strong data privacy protection and the elimination of undue biomedical research restrictions can be achieved on a global scale.

24.5 Towards Better Calibration of Biomedical Research, Health Service Delivery, and Privacy Protection

This chapter has illustrated that existing data privacy protection frameworks may inadequately resolve the tension between facilitating research and medical care through data sharing while also protecting against data misuse and privacy threats. Many legal frameworks were constructed before the emergence of Big Data analytics, eHealth, and commonplace cross-border collaboration. Rather than protecting privacy and advancing research and health, they may now be creating unnecessary barriers to sharing medical data across borders, and consequently impeding scientific discovery, while also leaving important data privacy protection gaps. As noted in an influential 2009 US Institute of Medicine Report: “If society seeks to derive the benefits of medical research in the form of improved health and health care, information should be shared to achieve that greater good, and governing regulations should support the use of such information, with appropriate oversight” [40]. In this section, we identify pathways that are available to better calibrate these three societal pillars of biomedical research, health service delivery, and privacy protection.

First, to achieve greater harmonisation of currently disjointed data privacy laws around the world so as to promote more efficient and responsible sharing of medical data, we may need to focus less on working towards a common framework of prescriptive data privacy rules and more on developing foundational responsible

data sharing principles, harkening back to an approach like that of the OECD Privacy Guidelines. As Professor Weber observes:

Due to social, historical, and cultural differences . . . harmonisation [of data protection standards] will remain at a high level of abstraction and at a low level as far as substance is concerned. The difficulties in agreeing on the form of the legal framework, in selecting the standards on which such an instrument would be based, in determining the scope of the instrument, and in agreeing on an international organisation to coordinate the work are too substantial for a high harmonisation level to be achieved soon [75].

Medical legal scholars Graeme Laurie and Nayha Sethi propose that “principles-based regulation” (PBR), as opposed to “rules-based regulation” offers less strict pre- and prescriptive rules for framing approaches to governance and decision-making of health-related research using personal data. As they state, PBR is envisaged “as the use of broadly-stated objectives, standards and values by which individuals and institutions should conduct themselves when using data for research purposes” [46].

Building on this concept, and extrapolating from the medical data legal and policy instruments covered in this chapter, we posit the following principles as the starting point of a principles-based approach to regulating the sharing medical data around the globe in an interoperable and responsible manner:

Transparency: The data privacy regulatory system should be transparent: individuals should be able to easily understand and access information about the collection, use, and disclosure of their medical data and privacy and security practices. In the context of genomic data, thought should be given to providing a mechanism for data access to blood relatives.

Control: Individuals should be able to exercise a reasonable degree of control over their own medical data and how they are collected, used, and disclosed, with an understanding that the scale and interconnectedness of medical data and biomedical research may limit claims to a “right” to (fully) control the data.

Quality: Medical data should be accurate, complete, and kept up-to-date to the extent necessary for collection purposes.

Security: Medical data should be accompanied by adequate security safeguards with regard to the risk involved and the nature and uses of the medical data.

Proportionality: Specific regulatory obligations imposed on data controllers should be proportionate to the reasonable likelihood of benefits arising from medical data sharing, as well as severity of the harm posed by the processing of medical data.

Flexibility: Regulations must be adaptable to changing norms, standards, innovations, and other regulations.

Evidence-basis: Regulations should be based on foreseeable use of potentially accessible, available, and valid data, and on current data privacy, medical, and technological scholarship, as well as foreseeable developments on the horizon in those fields.

Accountability: Both medical-data controllers and data processors should be responsible and should be held accountable for unlawful data processing downstream.

Risk Assessment: For any new collection, use, or disclosure of medical data, privacy risks and strategies to mitigate them should be identified and assessed.

Transborder Flows: Sharing of medical data to other jurisdictions that offer at least comparable safeguards for the protection of medical data should not be obstructed on the sole basis that the data will enter a new jurisdiction.

The benefit of a PBR approach is that it can allow any given “decision-maker to reflect on broad-based values and commonly-agreed objectives to determine through deliberation and reflection what action best fits in accordance with the particular value(s) advanced, avoiding reliance upon detailed anticipatory drafting for every conceivable situation” [46]. The principles identified above, coupled with instances of best practices (i.e., examples of “principles in action”), could help those processing medical data to navigate the often impenetrable, obstructing thickets of data privacy regulation.

Regulators should consider adopting a proportionate approach to data privacy protection. Anonymised or de-identified medical data, and potentially pseudonymised medical data, should be subject to less restrictive legislative provisions (and thus achieve some form of legislative exemption or reduction in regulatory burden) in the contexts and following the processes for which evidence-based research supports their use. Privacy regulation should be continuously updated in light of evolving methods and technologies to ensure that re-identification risks remain remote.

At the same time, organisations using medical data should ensure that their systems adequately protect privacy, by developing a medical data privacy plan which, at a minimum:

- Identifies all people who will have access to the data, and proportionately controls access to the data (including through data-use agreements that require recipients of anonymised or pseudonymised medical data to abide by a set of privacy-promoting conditions).
- Identifies a person as responsible for maintaining data protection safeguards.
- Establishes measures to prevent and sanction the deliberate re-identification of individuals from medical data that have had direct identifiers removed (absent explicit consent from participants or express legal authorisation).
- Describes the measures for protecting the physical, software, and remote-server security of the data.
- Prevents unauthorised or unauthenticated people from accessing the medical data through the use of e.g., firewalls, data encryption, and robust password protection schemes.
- Assesses the risk of re-identification (if medical data are anonymised or pseudonymised) when research studies are planned or medical data will be shared, and reviews this risk regularly during the lifetime of the study.
- Provides a contingency plan for dealing with any breach of privacy or confidentiality.

24.6 Conclusion

Biomedical research, health-service delivery, and data privacy mutually promote one another in the aim of arriving at a common purpose: to benefit individuals and society. Privacy controls help individuals flourish as members of society, including by establishing the conditions necessary for willing participation in research and the healthcare system. Research and healthcare, in turn, contribute to better health and community wellbeing. Both research participants and patients expect and deserve robust privacy protection of their medical data. As biomedical research, science, and medicine advance, and as more medical data are collected, shared, and used for myriad purposes, protecting privacy and maintaining confidentiality are becoming increasingly complex but vital tasks.

Researchers, clinicians, and regulators alike should be mindful of emerging technologies that sufficiently protect (or threaten) patient and participant privacy [43, 74]. Medical data from biobanks and cohort studies are increasingly shared within and across institutions (and borders) to create combined datasets which can be queried to ask complex scientific questions with due regard for data privacy principles [55]. In this environment, technologies will increasingly be relied upon to uphold standards of protection for the sharing of medical data. But they must be combined with robust scientific, ethics, and data access governance systems that anticipate and address the sharing of data with other researchers or healthcare professionals across jurisdictions, and that explain that even if data are de-identified, there always remains a residual risk of re-identification, though the regulatory goal is *acceptable* risk [23]. Technologies must offer flexible means of processing and widely sharing medical data while protecting privacy in accordance with applicable legislation and policies.

Ultimately, if we are to globally harness the power of medical data safely and securely to advance the wellbeing of individuals and society, it is critical that regulators, organisations, and individuals alike recognize the benefits that accrue from medical data sharing. It is equally imperative, however, that we create interoperable data privacy governance frameworks and systems that are robust yet flexible so that they are adaptable to new technologies and models of research and health service delivery.

References

1. Academy of Medical Sciences: Personal data for public good: using health information in medical research. <http://www.acmedsci.ac.uk/policy/policy-projects/personal-data/> (2006). Accessed 22 June 2015
2. Agaku, I.T., Adisa, A.O., Ayo-Yusuf, O.A., Connolly, G.N.: Concern about security and privacy, and perceived control over collection and use of health information are related to withholding of health information from healthcare providers. J. Am. Med. Inform. Assoc. **21**, 374–378 (2014)

3. Arias, J.J., G, G.P.K., Campbell, E.G.: The growth and gaps of genetic data sharing policies in the united states. *J. Law Biosci.* **2**, 56–58 (2015)
4. Article 29 Data Protection Working Party: Opinion 15/2011 on the definition of consent. http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2011/wp187_en.pdf (2011). Accessed 22 June 2015
5. Article 29 Data Protection Working Party: Opinion 05/2014 on anonymisation techniques. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (2014). Accessed 22 June 2015
6. Article 29 Data Protection Working Party: Letter from article 29 working party to paul timmers, director of sustainable and secure society, directorate, dg connect, regarding health data in apps and devices (5 february 2015). http://ec.europa.eu/justice/data-protection/article-29/documentation/other-document/files/2015/20150205_letter_art29wp_ec_health_data_after_plenary_annex_en.pdf (2015). Accessed 22 June 2015
7. BC IPC (British Columbia Office of the Information & Privacy Commissioner): A prescription for legislative reform: improving privacy protection in BC's health sector. <https://www.oipc.bc.ca/special-reports/1634> (2014). Accessed 22 June 2015
8. Beyleveld, D., Townend, D., Rouille-Mirza, S., Wright, J.: The Data Protective Directive and Medical Research Across Europe. Ashgate, Aldershot (2005)
9. Boniface, M.A.: Privacy and Data Protection in Africa. Scholars Press, Saarbrucken (2014)
10. Bygrave, L.A.: Data Privacy Law: An International Perspective. Oxford University Press, Oxford (2014)
11. Bygrave, L.A.: Information concepts in law: generic dreams and definitional daylight. *Oxf. J. Leg. Stud.* **35**, 91–120 (2015)
12. Canada: 1983 privacy act. <http://laws-lois.justice.gc.ca/eng/acts/P-21> (1983). Accessed 22 June 2015
13. Canada: Personal information protection and electronic documents act. <http://laws-lois.justice.gc.ca/eng/acts/P-8.6> (2000). Accessed 22 June 2015
14. Canadian Institutes of Health Research: Cihr open access policy. <http://cihr-irsc.gc.ca/e/46068.html> (2013). Accessed 22 June 2015
15. Cavoukian, A., Emam, K.E.: De-identification protocols: essential for protecting privacy. http://www.privacybydesign.ca/content/uploads/2014/06/pbd-de-identification_essential.pdf (2014). Accessed 22 June 2015
16. Contreras, J.L.: NIH's genomic data sharing policy: timing and tradeoffs. *Trends Genet.* **31**, 55–57 (2015)
17. Council of Canadian Academies: Accessing health and health-related data in Canada. <http://www.scienceadvice.ca/en/assessments/completed/health-data.aspx> (2015). Accessed 22 June 2015
18. Council of Europe: Convention for the protection of individuals with regard to automatic processing of personal data. <http://conventions.coe.int/Treaty/en/Treaties/Html/108.htm> (1981). Accessed 22 June 2015
19. Council of Europe: Recommendation no. r (97) 5 of the committee of ministers to member states on the protection of medical data. <http://wcd.coe.int/ViewDoc.jsp?id=571075> (1997). Accessed 22 June 2015
20. Council of Europe: Additional protocol to the convention for the protection of individuals with regard to automatic processing of personal data regarding supervisory authorities and transborder data flows. <http://conventions.coe.int/Treaty/en/Treaties/HTML/181.htm> (2001). Accessed 22 June 2015
21. Council of Europe: Consultative committee of the convention for the protection of individuals with regard to automatic processing of personal data [ets no. 108]: proposals of modernisation. [http://www.coe.int/t/dghl/standardsetting/dataprotection/TPD_documents/T-PD\(2012\)4Rev3E%20-%20Modernisation%20of%20Convention%20108.pdf](http://www.coe.int/t/dghl/standardsetting/dataprotection/TPD_documents/T-PD(2012)4Rev3E%20-%20Modernisation%20of%20Convention%20108.pdf) (2012). Accessed 22 June 2015
22. DeCew, J.: In Pursuit of Privacy: Law, Ethics, and the Rise of Technology. Cornell University Press, Ithaca (1997)

23. Emam, K.E., Alvarez, C.: A critical appraisal of the article 29 working party opinion 05/2014 on data anonymisation techniques. *Int. Data Priv. Law* **5**, 73–87 (2015)
24. Emam, K.E., Jonker, E., Ar buckle, L., Malin, B.: A systematic review of re-identification attacks on health data. *PLoS One* **6** (2011)
25. European Commission: Proposal for a regulation of the european parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (general data protection regulation). http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf (2012). Accessed 22 June 2015
26. European Commission: Proposal for a regulation of the european parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (general data protection regulation) - preparation of a general approach. <http://data.consilium.europa.eu/doc/document/ST-9565-2015-INIT/en/pdf> (2015). Accessed 22 June 2015
27. European Parliament: Committee on civil liberties, justice and home affairs draft report on the proposal for a regulation of the european parliament and of the council on the protection of individual with regard to the processing of personal data and on the free movement of such data (general data protection regulation). http://www.europarl.europa.eu/meetdocs/2009_2014/documents/libe/pr/922/922387/922387en.pdf (2012). Accessed 22 June 2015
28. European Union: Directive 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML> (1995). Accessed 22 June 2015
29. European Union: Charter of fundamental rights of the european union. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2010:083:0389:0403:en:PDF> (2010). Accessed 22 June 2015
30. Expert Advisory Group on Data Access: Statement for EAGDA funders on re-identification. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp055972.pdf (2013). Accessed 22 June 2015
31. Genome Canada: Data release and resource sharing. <http://genomecanada.ca/medias/PDF/EN/DataReleaseandResourceSharingPolicy.pdf> (2008). Accessed 22 June 2015
32. Government of Canada: Tri-agency open access policy on publications. <http://www.science.gc.ca/default.asp?lang=En&n=F6765465-1> (2015). Accessed 22 June 2015
33. Greenleaf, G.: Global data privacy laws: 89 countries, and accelerating, queen mary university of London, school of law legal studies research paper no. 98/2012. <http://ssrn.com/abstract=2000034> (2012). Accessed 22 June 2015
34. Greenleaf, G.: Asian Data Privacy Laws: Trade & Human Rights Perspectives. Oxford University Press, Oxford (2014)
35. Greenleaf, G.: Global data privacy laws 2015: 109 countries, with european laws now a minority. *Priv. Laws Bus. Int. Rep.* **133**, 18–28 (2015)
36. Hallinan, D., Friedewald, M.: Open consent, biobanking and data protection law: can open consent be ‘informed’ under the forthcoming data protection regulation? *Life Sci. Soc. Policy* **11**, 1 (2015)
37. HEW (US Department of Health, Education and Welfare): Records, computers and the rights of citizens: report of the secretary’s advisory committee on automated personal data systems. <http://www.justice.gov/sites/default/files/opcl/docs/rec-com-rights.pdf> (1973). Accessed 22 June 2015
38. Homer, N. et al.: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet.* **4**, e1000167 (2008)
39. ILRDP Kantor Ltd: Comparative study on different approaches to new privacy challenges, in particular in the light of technological developments. http://ec.europa.eu/justice/policies/privacy/docs/studies/new_privacy_challenges/final_report_en.pdf (2010). Accessed 22 June 2015
40. Institute of Medicine: Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. National Academies, Washington (2009)

41. International Conference of Data Protection and Privacy Commissioners: International standards on the protection of personal data and privacy: the madrid resolution. http://www.privacycommission.be/sites/privacycommission/files/documents/international_standards_madrid_2009.pdf (2009). Accessed 22 June 2015
42. Kenyon, A.T., Richardson, M.: New Dimensions in Privacy: International and Comparative Perspectives. Cambridge University Press, Cambridge (2010)
43. Knoppers, B.M., Dove, E.S., Litton, J.E., Nietfeld, J.J.: Questioning the limits of genomic privacy. *Am. J. Hum. Genet.* **91**, 577–578 (2012)
44. Knoppers, B.M., Saginur, M.: The babel of genetic data terminology. *Nat. Biotechnol.* **23**, 925–927 (2005)
45. Kuner, C.: Transborder Data Flows and Data Privacy Law. Cambridge University Press, Oxford (2013)
46. Laurie, G., Sethi, N.: Towards principles-based approaches to governance of health-related research using personal data. *Eur. J. Risk Regul.* **4**, 43–57 (2013)
47. Lowrance, W.W.: Privacy, Confidentiality, and Health Research. Cambridge University Press, Oxford (2012)
48. Moraia, L.B. et al.: A comparative analysis of the requirements for the use of data in biobanks based in finland, germany, the netherlands, norway and the united kingdom. *Med. Law Int.* **14**, 187–212 (2014)
49. National Institutes of Health: Policy for genome-wide association studies. <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html> (2007). Accessed 22 June 2015
50. National Institutes of Health: Modifications to genome-wide association studies (GWAS) data access. <https://gds.nih.gov/pdf/Data%20Sharing%20Policy%20Modifications.pdf> (2008). Accessed 22 June 2015
51. National Institutes of Health: NIH genomic data sharing policy. http://gds.nih.gov/PDF/NIH_GDS_Policy.pdf (2014). Accessed 22 June 2015
52. National Institutes of Health: Supplemental information to the national institutes of health genomic data sharing policy. http://gds.nih.gov/PDF/Supplemental_Info_GDS_Policy.pdf (2014). Accessed 22 June 2015
53. NIH-DOE Joint Subcommittee: NIH-DOE guidelines for access to mapping and sequencing data and material resources (adopted 7 December). <http://www.genome.gov/10000925> (1992). Accessed 22 June 2015
54. Nissenbaum, H.: Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford University Press, Stanford (2010)
55. Nuffield Council on Bioethics: The collection, linking and use of data in biomedical research and health care: ethical issues. http://nuffieldbioethics.org/wp-content/uploads/Biological_and_health_data_web.pdf (2015). Accessed 22 June 2015
56. OECD: The OECD privacy framework. http://oecd.org/sti/ieconomy/oecd_privacy_framework.pdf (2013). Accessed 22 June 2015
57. O'Neill, O.: Some limits of informed consent. *J. Med. Ethics* **4** (2003)
58. Phoenix SPI: Survey of canadians on privacy-related issues. final report. https://www.priv.gc.ca/information/por-rop/2013/por_2013_01_e.asp (2013). Accessed 22 June 2015
59. Power, M.: The Law of Privacy. LexisNexis Canada, Markham (2013)
60. Smith, R., Shao, J.: Privacy and e-commerce: a consumer-centric perspective. *Electron. Commer. Res.* **7**, 89–116 (2007)
61. Solove, D.J., Schwartz, P.M.: Information Privacy Law, 5th edn. Wolters Kluwer, New York (2015)
62. Taylor, M.: Genetic Data and the Law: A Critical Perspective on Privacy Protection. Cambridge University Press, Cambridge (2012)
63. Tene, O.: Privacy law's midlife crisis: a critical assessment of the second wave of global privacy laws. *Ohio State Law J.* **74**, 1217–1261 (2013)
64. Tzanou, M.: Data protection as a fundamental right next to privacy? 'reconstructing' a not so new right. *Int. Data Priv. Law* **3**, 88–99 (2013)

65. United Kingdom: Data protection act 1998. <http://legislation.gov.uk/ukpga/1998/29> (1998). Accessed 22 June 2015
66. United Kingdom: The data protection (processing of sensitive personal data) order 2000. <http://www.legislation.gov.uk/ksi/2000/417/schedule/made> (2000). Accessed 22 June 2015
67. United Nations: General assembly resolution 2450 of 19 December 1968. Doc E/CN.4/1025 (1968)
68. United Nations: Points for possible inclusion in draft international standards for the protection of the rights of the individual against threats arising from the use of computerized personal data systems. Doc E/CN.4/1233 (1976)
69. United Nations: Guidelines concerning computerized personal data files (UN general assembly resolution 45/95 of 13 December 1990). Doc E/CN.4/1990/72 (1990)
70. United States: Code of federal regulations. title 45: public welfare. part 160: general administrative requirements. http://www.ecfr.gov/cgi-bin/text-idx?tpl=/ecfrbrowse/Title45/45cfr160_main_02.tpl (2014). Accessed 22 June 2015
71. United States: Code of federal regulations. title 45: public welfare. part 164: security and privacy. http://www.ecfr.gov/cgi-bin/text-idx?tpl=/ecfrbrowse/Title45/45cfr164_main_02.tpl (2014). Accessed 22 June 2015
72. United States Department of Commerce: Safe harbor privacy principles. http://www.export.gov/safeharbor/eu/eg_main_018475.asp (2000). Accessed 22 June 2015
73. US Privacy Protection Study Commission: Personal privacy in an information society. US Government Printing Office, Washington (1977)
74. Wallace, S.E., Gaye, A., Shoush, O., Burton, P.R.: Protecting personal data in epidemiological research: DataSHIELD and UK law. *Public Health Genomics* **17**, 149–157 (2014)
75. Weber, R.H.: Transborder data transfers: concepts, regulatory approaches and new legislative initiatives. *Int. Data Priv. Law* **3**, 117–130 (2013)
76. Wellcome Trust: Policy on data management and sharing. <http://www.wellcome.ac.uk/about-us/policy/policy-and-position-statements/wtx035043.htm> (2010). Accessed 22 June 2015
77. Wellcome Trust: Summary report of qualitative research into public attitudes to personal data and linking personal data. <http://www.wellcome.ac.uk/About-us/Publications/Reports/Public-engagement/WTP053206.htm> (2013). Accessed 22 June 2015
78. World Health Organisation: Legal frameworks for ehealth: based on the findings of the second global survey on eHealth. http://whqlibdoc.who.int/publications/2012/9789241503143_eng.pdf (2012). Accessed 22 June 2015
79. Younger Committee: Report of the committee on privacy. Home Office, Cmnd 5012. HMSO, London (1972)

Chapter 25

HIPAA and Human Error: The Role of Enhanced Situation Awareness in Protecting Health Information

Divakaran Liginlal

Abstract Several contemporary studies have identified human error as a major cause of privacy breaches in healthcare organizations. In this chapter, we first highlight the costs healthcare organizations incur from HIPAA privacy breaches. We then discuss the concept of situation awareness (SA) and its link with privacy protection. Situation awareness represents individuals' awareness of what is happening in their surroundings and their understanding of how information, events, and actions affect their goals and objectives. Applying Endsley's three-level SA framework helps us to identify specific types of SA errors and build scenarios of privacy breaches arising from SA errors. Using a taxonomy of SA errors derived from Endsley's work, we analyzed the 21 cases of HIPAA privacy breaches in which the Office for Civil Rights has reached a resolution agreement. The results bring into focus the often neglected situational aspects of privacy protection and help to better understand the latent causes of privacy breaches along with their related implications for policy formulation, system design, and user training.

25.1 Introduction

Since Hippocrates' times, the importance of safeguarding a patient's health information has been enshrined in the Hippocratic Oath: "and whatsoever I shall see or hear in the course of my profession, as well as outside my profession in my intercourse with men, if it be what should not be published abroad, I will never divulge, holding such things to be holy secret." Today, the need to protect individuals' health information is of interest not only from the perspective of personal privacy but also because of the potentially adverse effects on their employment status and insurance coverage.

D. Liginlal (✉)

Dietrich College of Humanities and Social Sciences, Carnegie Mellon University,
Pittsburgh, PA, USA

e-mail: liginlal@cmu.edu

With the rapid evolution of healthcare toward managed care and the accelerated adoption of electronic health record (EHR) systems, information exchange networks, and health insurance exchanges, patients face a higher risk of loss of privacy of their health information. Particularly, the rise in the use of Web technologies, personal health monitoring systems, physician order entry systems, and a variety of other clinical information systems has exacerbated the potential for privacy breaches along with the associated costs they impose on healthcare organizations. Legislation such as the Health Insurance Portability and Accountability Act (HIPAA) [26] and the Health Information Technology for Economic and Clinical Health (HITECH) [10] aims to ensure that healthcare organizations safeguard patients' health information against security breaches and other forms of disclosure [2].

The need to protect a patient's health information against deliberate or inadvertent misuse or disclosure is enshrined in the HIPAA Privacy Rule. "Individually identifiable health information" is defined as information, including demographic data that relates to an individual's past, present, or future physical or mental health or condition, the provision of healthcare to an individual and the associated payment information, and other information that specifically identifies an individual. A healthcare provider, a health plan, a clearinghouse, and any healthcare provider who transmits health information in electronic form in connection with transactions is considered a covered entity under HIPAA. The term "protected health information" (PHI) refers to any such information held or transmitted by a covered entity or its business associates, in any form or media, whether electronic, paper, or oral. The Privacy Rule upholds the rights of patients with respect to PHI held by covered entities and their business associates. At the same time it balances the need for disclosure of required information in the course of delivering quality care. The Privacy Rule mandates that covered entities must use or disclose the minimum necessary PHI for a specific purpose [9]. In addition, the HIPAA Security Rule specifies a series of safeguards—technical, administrative, and physical—to assure the confidentiality, integrity, and availability [12] of electronic protected health information (ePHI). Unfortunately, even with sufficient safeguards, the reported incidents of privacy breaches in healthcare organizations and the resulting loss of personal health information have been on the rise [16].

Consistent with the Privacy Rule, patients have the right to access their health information and also to know how covered entities use and disclose that information. Covered entities must obtain written authorization for the use and disclosure of such information except as it is needed in providing care, payment, and other healthcare operations. Further, when violation of these rights occur, patients have the right to lodge a formal complaint with a covered provider or health plan, or with the U.S. Department of Health and Human Services (HHS) Office for Civil Rights (OCR) and obtain redress, often resulting in civil monetary and criminal penalties for the covered entity. By enforcing the Privacy and Security Rules, the OCR helps to protect the privacy of individuals' health information held by covered entities. Some of the related providers and insurers may include: doctors and nurses, pharmacies, hospitals, clinics, nursing homes, health insurance companies,

health maintenance organizations (HMOs), employer group health plans, and certain government programs that pay for health care, such as Medicare and Medicaid [26].

For the purposes of this chapter and within the context of healthcare, we associate the term “privacy protection” with the actions taken by a person (e.g., a healthcare worker or practitioner) working for a covered entity or business associate to protect the privacy and security of an individual’s health information. We consider privacy protection as a nonmonotonic, dialectic, and dynamic process in which the environmental and situational contexts play an important role [9, 11, 14, 23, 27]. The construct “situation awareness” (SA) represents individuals’ awareness of what is happening in their surroundings and their understanding of how information, events, and actions affect their goals and objectives [8]. Individuals with high SA levels are likely to understand and predict a situation more accurately and control their privacy protection behavior accordingly. The major contribution of this chapter is this consideration of SA in protecting the privacy of health information—in other words, the interplay between the individual, situation, and context in a healthcare setting.

Before delving into the role of SA, one needs to understand on a global scale the impact of privacy breaches within the healthcare sector vis-à-vis other industry sectors and the corresponding role of human error. The IBM and Ponemon Institute [19] study of 350 companies on the cost of data breaches in 11 countries, provides supporting information in this regard. The highest average per capita cost is incurred in the U.S.A and Germany. Although the average global cost of a data breach per lost record is only about \$150 across different industry sectors, the highest average cost of \$363 is incurred in healthcare organizations. Also, the loss of customers increases the cost of data breaches and such churn is estimated to be highest within the health and pharmaceutical industries. Most importantly, human error-related average breach cost of \$134 remains significant and not too different than the average cost of \$170 for malicious or criminal acts, which often results from poorly configured systems or other human errors. The IBM and Ponemon study also uncovers significant variations in the likelihood of a data breach involving a minimum of 10,000 records across the 11 countries surveyed. While Canada and Germany have the least likelihood of a data breach, countries such as Brazil and France have the highest chance of a data breach.

In capturing the situated aspects of privacy protection, we address a major gap in the extant privacy research. Lack of SA constitutes an important cause of human error in decision-making situations, which is the foremost, but often ignored, cause of privacy breaches [3, 15, 18, 22]. One of the primary goals of HIPAA is to streamline and reduce the cost of delivering healthcare. Regardless, the implementation of HIPAA Privacy and Security Rule compliance are likely to result in high initial costs to a healthcare organization. However, these high initial costs are likely to be much less than the potential damage to reputation and the salvage costs of recovering from a HIPAA-related privacy breach.

The results of this study have several practical implications for both healthcare organizations and healthcare workers who handle PHI. The chapter is especially relevant to healthcare systems and solutions providers, because it suggests means to reduce incidents of privacy breaches arising from lack of SA of system users and

operators. Equally important, the chapter provides insights into the importance of designing privacy-enhancing user interfaces that incorporate powerful help features and privacy alerts; it also recommends that organizations create effective employee training programs that specifically address enhanced SA.

The remainder of this chapter is organized as follows. In Sect. 25.2, we consider the role of human error in privacy breaches in healthcare organizations and the related costs. Then in Sect. 25.3, we define SA and apply the [8] three-level SA model and epidemiological framework to understand the importance of SA in privacy protection. We illustrate how SA errors lead to privacy breaches through an analysis of 21 cases of HIPAA privacy breaches in which the OCR has reached a resolution agreement. Finally, in Sect. 25.4, we present the suggested SA enhancement methods and propose avenues for future studies on this topic.

25.2 HIPAA, Privacy Breaches, and Related Costs

Since the advent of the HITECH rule, requiring organizations to notify HHS of the discovery of a data breach, the HHS has reported the compromise of the PHI of more than 120 million individuals! [25]. Details of the 12 largest notified privacy breaches to date, compiled from the OCR portal [24] are shown in Table 25.1.

For the organizations involved, besides damage to their reputation, the resulting tangible costs involved lawsuits, fines, and the need to provide free credit monitoring and identity theft protection for affected individuals for a significant period of time. Two notable cases are Sutter, which has been hit with multiple lawsuits that could amount to between \$944 million and \$4.25 billion, and BlueCross BlueShield of Tennessee, which paid \$1.5 million in fines to the HHS and nearly \$17 million for protection, investigation, and member notification [17]. A majority of the cases involved accidental disclosure, loss of devices containing PHI, and theft because of improper physical safeguards. Analyzing the primary causes of these breaches reveals that most of them arise from lost or stolen devices that were not protected with the safeguards mandated by the Security Rule.

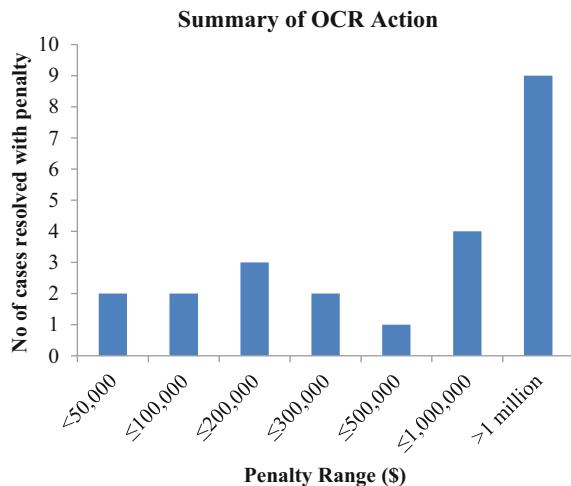
Since 2008, the OCR has concluded resolution agreements with 21 covered entities. A resolution agreement, according to the OCR, is a contract signed by HHS and a covered entity in which the covered entity agrees to perform certain obligations (e.g., staff training) and make reports to HHS, generally for three years. During this period, HHS monitors the covered entity's compliance with its obligations. A resolution agreement likely would include the payment of a resolution amount [25]. Examining Fig. 25.1, which shows the distribution of penalties imposed by OCR on these 21 covered entities, helps us to conclude that more than half the cases involved payments of \$1 million or more. These penalties are above and beyond other costs incurred in salvage operations.

Five other notable breach examples reveal interesting insights into the causal factors and the costs arising from a breach:

Table 25.1 The 12 largest privacy breaches identified by the OCR

Date reported	Covered entity	Description	Individuals affected	Primary cause
Mar 2015	Anthem	Hacker attack	79 million	Security vulnerability
Mar 2015	Premera Blue Cross	Hacker attack	11 million	Security vulnerability
Oct 2011	Science Applications International Co.	Backup tapes lost	4.9 million	Human error leading to loss
Aug 2014	Community Health Systems	Hacker attack	4.5 million	Security vulnerability
Aug 2013	Advocate Medical Group	Theft of four unencrypted computers	4.03 million	Human error leading to theft
Sep 2014	Xerox State Healthcare	Unauthorized access	2 million	Human error leading to disclosure
Apr 2011	Health Net Inc.	Lost nine server drives	1.9 million	Human error leading to loss
Feb 2011	North Bronx Healthcare Network	Stolen backup tapes	1.7 million	Human error leading to theft
Jun 2010	AvMed Inc.	Stolen unencrypted laptops	1.22 million	Human error leading to theft
Jul 2013	Montana Dept. Of Public Health and Human Services	Hackers gained server access for one year	1.06 million	Security vulnerability
Oct 2011	The Nemours Foundation	Missing unencrypted backup tapes	1.06 million	Human error leading to loss
Nov 2009	BlueCross BlueShield of Tennessee	Stolen unencrypted computer hard drives	1.02 million	Human error leading to theft

Fig. 25.1 Summary of OCR action on covered entities since 2008



1. The Beth Israel Deaconess Medical Center in Boston was fined \$100,000 by the state of Massachusetts. A physician failed to encrypt the PHI on his laptop, which was stolen. The fine covered a civil penalty, legal fees, and cost of instituting a HIPAA awareness program.
2. Triple-S Management Corp., an insurance holding company, was fined \$6.8 million for improperly handling the medical records of some 70,000 individuals. In this particular case, the covered entity reportedly mailed letters to its Medicare Advantage patients with the Medicare numbers visible from the exterior of the envelope.
3. A photocopier originally leased to Affinity Health Plan, a New York-based managed care plan, was later sold to a television network. The photocopier's hard disk contained medical information on nearly 350,000 individuals. The company had to pay \$1,215,780 to HHS for this violation, which was caused by overlooking the need to erase data before returning equipment to a lessee.
4. Inadequate physical security measures resulted in the theft of four unencrypted company computers from the premises of Advocate Medical Group, resulting in the compromise of the PHI of more than four million patients.
5. Microfilms belonging to Texas Health Harris Methodist Hospital that were given to Shred-it for destruction ended up in various public locations. The loss of the PHI on the microfilms resulted in the need to notify 277,000 patients of the potential breach.

Healthcare IT News (<http://www.healthcareitnews.com>) carries several more examples of such breaches. As categorized by [15], the underlying causal factors for a great majority of the examples illustrated in this chapter represent human error and arise from either mistakes or slips [20].

25.3 Situation Awareness and Privacy Protection

In this section, we first define the SA construct, then develop its theoretical underpinnings and its link to privacy protection. We examine how Endsley's three-level SA framework helps identify specific types of SA errors. Applying a taxonomy of human error derived from Endsley's work, we analyze prominent examples of HIPAA privacy breaches in which the Office for Civil Rights has reached a resolution agreement.

25.3.1 *Definition of Situation Awareness*

SA has been widely used as a construct in the aviation and engineering disciplines in which quick decisions in situation-specific settings are required and the consequences of such decisions can be devastating [8]. As its name implies, SA represents individuals' awareness of what is happening around them and their understanding of how information, events, and actions can affect their goals or objectives. Although several approaches exist for defining and measuring SA, a common understanding of SA relates to the situational factors underlying the divide between a human operator's understanding of the status of a system and its actual status and the need to bridge this gulf. SA is an action-level, task-oriented concept. It is affected by knowledge, skills, perception, concerns, experience, and other factors previously acquired by individuals. More important, it serves as a link between these prior factors and actual behavior rather than as an aggregation and abstraction of such factors.

One of the most widely accepted streams of SA research draws on an information-processing model developed by Endsley in [8]. In his approach, traditional psychological constructs such as attention, long-term memory, schemata, automaticity, and expectation are important elements that affect SA. By associating individual factors, environmental factors, and their interactions with SA, Endsley's model explains individuals' differences in SA. Endsley states that working memory capacity and limited attention restrict a person's SA and that a higher level of SA can be achieved by developing long-term memory through schemata and mental models. Pattern matching between critical cues in the environment helps achieve higher SA under conditions of uncertainty and incomplete information. Personal goals and objectives are important factors that determine how attention is directed and how information is perceived and interpreted. Endsley's model postulates three levels of SA: Level 1 (perception of elements in the environment), Level 2 (comprehension of the current situation), and Level 3 (projection of future status). A higher level of SA helps in better decision-making and consequent action and is always desirable.

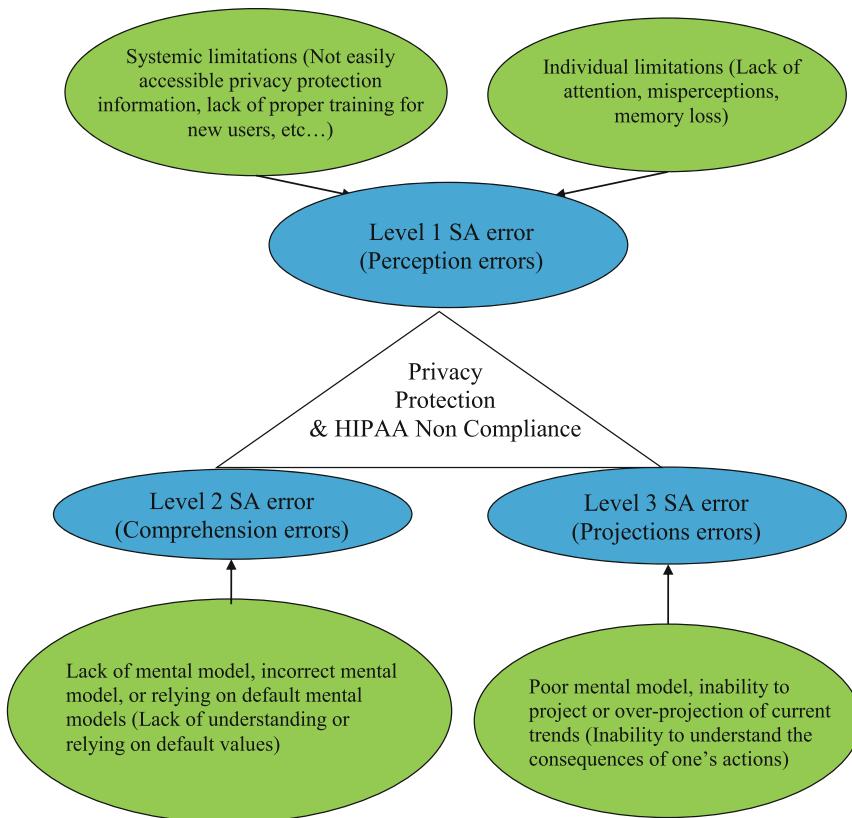


Fig. 25.2 Implications of Endsley's three-stage model on privacy protection

25.3.2 *Linking Situation Awareness to Privacy Breaches*

Human error is hard to predict and prevent and is the most frequent cause of breaches of privacy or security in organizations [3, 15, 18, 21]. We draw on Endsley's [8] epidemiological view of SA errors to identify such errors by level and subtype and then link these errors to information disclosure that leads to privacy loss. Figure 25.2 depicts how inadvertent disclosure of protected information because of a lack of SA leads to HIPAA privacy breaches.

25.3.2.1 **Level 1 SA: Failure to Correctly Perceive a Situation**

Users require clear and accurate directions to acquire accurate perceptions. Data availability is mostly related to the environment and system design. Contrary to the common belief that the cause of human errors leading to privacy breaches

may be attributed primarily to an individual, prior research suggests that the work environment is critical and that users are susceptible to errors because of systemic limitations [15]. The unavailability of data forces users to resort to guessing or to rely on default mental models, consequently leading to errors of judgment and faulty decisions. The presentation of data is as important as its availability. Critical data should be presented at the expected time and in the expected space to facilitate quick and accurate observation. Distractions need to be minimized, and careful attention should be given to effective design of system interfaces.

Users also commit SA errors by misreading the cues provided, which in turn leads to misperception. Misperception is often related to expectations [8]. When users find a cue at the expected time and place, they tend to assume that the information is correct without a careful second look. This type of misperception is rooted in psychology. The other type of misperception occurs when users simply misread the information because of physical disorientation and end up having poor Level 1 SA. To avoid observational failure or misperception, users must possess the skills to distinguish critical data from noncritical data and the ability to control their attention spans. User training, experience, and adequate monitoring may positively affect this process, and high workloads, understaffing, and multitasking may have negative effects. Finally, memory loss is another cause of SA perception errors. In some cases, users observe and perceive information correctly, but later forget it. This memory failure is related to individuals' personal ability, but it is also affected by lack of attention, overwork, and distractions.

25.3.2.2 Level 2 SA: Failure to Comprehend a Situation

In the comprehension stage of SA, errors may occur largely from three causes: *lack of a mental model, use of an incorrect mental model, or over-reliance on default mental models*. This stage is about how the significance and meaning of perceived information are assimilated. Endsley [8] argued that the lack of a mental model contributes to a low Level 2 SA. Individuals need a sound mental model to understand the meaning of perceived information. Although users may have a good knowledge of how the system functions, they may have developed habits of relying excessively on default mental models. This cause of SA errors is more frequently encountered in routine operations. Instead of integrating and comprehending perceived information according to a sound mental model, a user may choose to use his or her usual familiarity with the situation. In this case, familiarity and confidence overpower perceived information and knowledge.

25.3.2.3 Level 3 SA: Failure to Project a Situation into the Future

Level 3 SA errors result from a poor mental model and over projection of current trends. In some cases, users may have a poor model for projecting the future. This can occur even with accurate perception and good comprehension of the situation. In addition to the poor mental model, over-projection of current trends is another cause of SA error. In this case, users believe the current status will continue into

the future, although the perceived and comprehended situation says otherwise. This type of SA error in information disclosure defies rational explanation, but laziness, fatigue, or brashness may be contributing factors.

25.3.3 *SA and HIPAA Privacy Breaches*

Figure 25.3 contains a simplified illustration of the flow of information in a typical patient visit to a healthcare network.

Registration: As part of the registration process, a receptionist collects a patient's identifying information and links it to an existing medical record, or if none exists, creates one. The patient also receives a notice of privacy practices, as required by HIPAA, and acknowledges reading and accepting its terms. The collected information is stored in a clinical information system (indicated as EHR in Fig. 25.3).

Treatment: A designated medical worker, often a nurse, looks up key patient information, discusses and records the patient's health-related concerns/symptoms, and records the patient's vital signs. A doctor then looks up the health history of the patient, along with the reasons for the visit, and examines the patient. The doctor may order additional diagnostic tests. The clinical department also sends its billing information for the treatment so it can be forwarded to the insurance company.

Diagnostics: The patient visits a diagnostic center where the lab staff looks up relevant information and performs the required tests. The lab test results are sent back to the doctor to prescribe treatment.

Prescription: The doctor may prescribe medication for the patient. This information is often forwarded to the pharmacy. The patient visits the pharmacy, where a clerk at the counter looks up the patient's information and validates insurance information. The pharmacist looks up the prescription and prepares the medicines. The prescribed medication is then dispensed to the patient.

Payment: The insurance company receives the patient's treatment information from the hospital. The company assesses the patient's insurance coverage and determines the amount to pay the hospital and if necessary, bills the customer for any balance. Finally, the bill is mailed to the patient, and payment is made to the hospital.

Follow up: The hospital uses patient contact and medical information stored in the EHR to notify the patient of follow-up checkups or other important preventive care information. For compliance purposes this information is required to be kept for nearly a decade beyond the last date the patient was treated.

At every stage of this information flow, we can identify plausible breach scenarios involving loss of SA. Figure 25.4 for the registration phase illustrates how the three stages of SA can be linked to potential HIPAA privacy violations. In what follows, we discuss in more detail each of these three levels in the specific context of

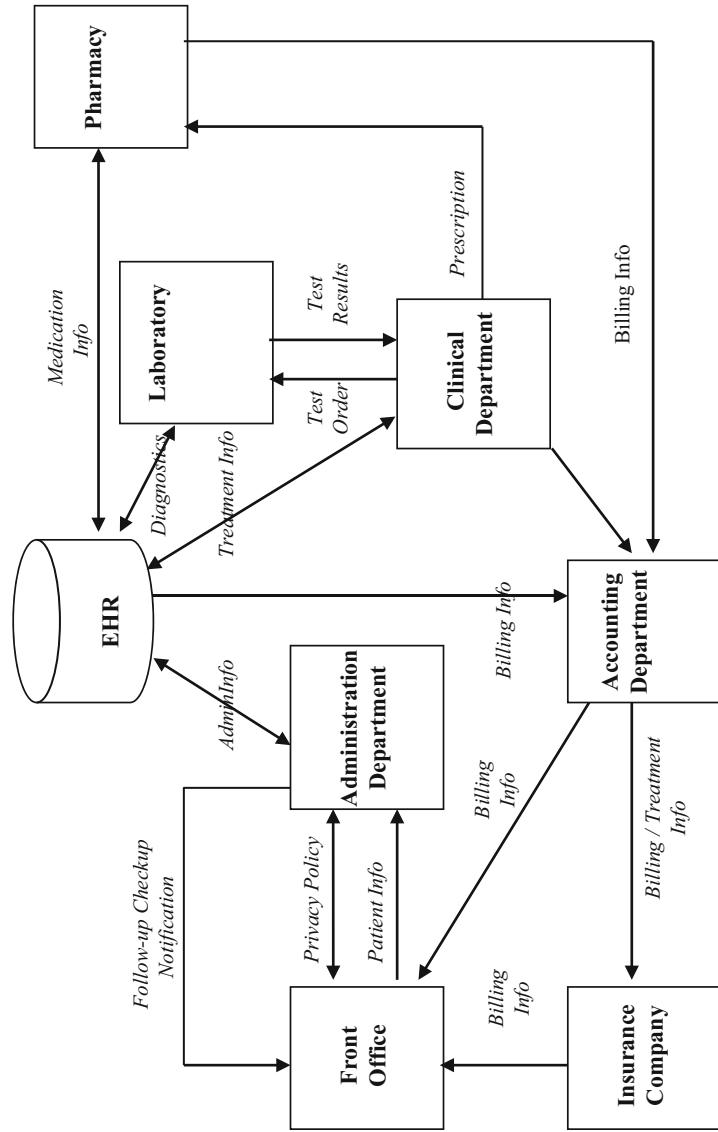


Fig. 25.3 A simplified illustration of information flow in a healthcare setting

Level 1 SA (Perception): A registration clerk took a phone call while she was retrieving a patient's PHI on a station that could be accessed by others. Without completing the registration, she left the information visible to others on the screen and walked away to attend to another task. This is an instance of failure to correctly perceive the situation through simple omission, memory loss, or a high task load that prevented her from observing the required privacy practice of logging out before leaving the station.

Level 2 SA (Comprehension): A registration clerk trashed a printout containing PHI without shredding it. This is a plausible case of faulty knowledge or over-reliance on defaults.

Level 3 SA (Projection): A registration clerk was busy and gave a temporary worker her access code to the system for the day. The temporary worker used the code to access medical information of people she knew. In this case, the temporary worker had not received proper HIPAA training, and the registration clerk's poor mental model resulted in her failure to project the potential consequences of her action.

Fig. 25.4 Example of SA errors in the registration process

Some examples are:

- Recording privacy policy agreement incorrectly because of unfamiliarity with instructions or an oversight (Registration stage).
- Failing to provide notice of privacy practices to a father or his minor daughter before a mental health evaluation (Registration stage).
- Displaying logbooks containing patients' prescription information in a manner that makes PHI visible to the public at a pharmacy counter (Prescription stage).
- Mistakenly placing a customer's insurance card in another customer's prescription bag (Prescription stage).
- Discussing medical status of a patient within earshot of other patients in a waiting room (All stages).
- Flagging medical records with a red sticker with the word "AIDS" on the outside cover and handling the records in a way that lets other patients and staff without a need to know read the sticker (All stages).
- Faxing a patient's medical records to the patient's place of employment instead of to the patient's new healthcare provider (Follow-up stage).

Fig. 25.5 Level 1 SA error—failure to correctly perceive a situation or misperception of information

HIPAA privacy breaches. A number of examples based on actual HIPAA breaches that the OCR has acted upon are highlighted in Figs. 25.5, 25.6, and 25.7 for each level of SA failure [25].

Some examples are:

- Trashing materials containing PHI without shredding (All stages).
- Accessing PHI of family members, friends, or colleagues (All stages).
- Emailing sensitive information with recipients in the CC field instead of the BCC field (Follow-up stage).
- Leaving a telephone message detailing a patient's condition with the daughter of the patient (Follow-up stage).
- Sending medical records to a disability insurance company without the patient's authorization (Payment stage).
- Incorrectly submitting a worker's compensation claim or sending PHI without authorization to a patient's employer (Payment stage).
- Failing to provide access to medical records (Registration stage).
- Charging a patient a fee for providing medical records (Registration stage).
- Hospital employee's supervisor accessing and disclosing employee's medical record for operational or other administrative purposes without the authorization of the employee (All stages).
- Disclosing a patient's medical information, without the prior written consent of the patient, to one of the following: (Follow-up stage)
 - public media as a public safety warning,
 - judicial body in response to a subpoena in certain cases, or
 - a research entity for recruitment purposes.
- Incorrectly requiring a patient to sign privacy agreements over and beyond what is guaranteed in the Privacy Rule (Registration stage).

Fig. 25.6 Level 2 SA error—failure to comprehend situation or improper comprehension of information

Some examples are:

- Giving escalated access rights to an assistant to allow completion of a task and not subsequently revoking the rights (All stages).
- Not purging PHI from a flash drive or other device (e.g., photocopier) that later became accessible to others.
- Accessing sites infected with malware from computers containing unencrypted PHI.
- Failure to update operating system or follow other safeguards such as encryption leading to vulnerability that is exploited.
- Using public computers to access PHI.

Fig. 25.7 Level 3 SA error—incorrect projection of situation into the future

25.3.3.1 Level 1 SA: Failure to Correctly Perceive a Situation

Level 1 SA errors may arise from a lack of availability of supporting information or when data is difficult to detect or perceive. The case of a covered entity failing to provide notice of privacy practices before a mental health evaluation is a specific example of this. However, when the required instructions and best practices are available and presented well, achieving a high level of perception becomes more dependent on the individuals themselves.

For healthcare employees, one of the most important pieces of information in the context of privacy is an awareness of their surroundings and who is privy to overhearing or seeing confidential information. Even if sufficient guidelines and training are provided, a user may commit an SA error through an oversight. Maintaining a high attention level is an important requirement for this task. Discussing the medical status of a patient within earshot of other patients in a waiting room or positioning a computer screen in such a way that unauthorized persons can see the data it displays are both examples of this category of SA errors. Poor perception often occurs when users are distracted by less important data and thereby fail to properly attend to critical ones. Some examples of Level 1 SA errors are shown in Fig. 25.5.

25.3.3.2 Level 2 SA: Failure to Comprehend a Situation

In many cases, even if information is correctly perceived, there may be a failure to comprehend its significance or meaning. Often this arises from a lack of an appropriate mental model for combining multiple pieces of information in association with pertinent goals [13]. Charging a patient a fee for providing medical records or disclosure of medical information to the judiciary or to public media, are examples of this. However, cases such as the failure to shred material containing PHI arises from over-reliance on default mental models. Some additional examples of Level 2 SA errors are shown in Fig. 25.6.

25.3.3.3 Level 3 SA: Failure to Project a Situation into the Future

Although in many cases individuals are fully cognizant of what they are doing, they have a poor model for projecting the consequences of their actions. Consider the case of an employee giving elevated access rights to an assistant so the latter could complete some pending tasks. This was done in good faith, but the employee's failure to revoke these rights so tasks could be delegated in the future exemplifies a Level 3 SA error because of the likelihood of the assistant abusing the privilege by viewing PHI that only her supervisor was authorized to access.

One may argue that the case of Affinity Health Plan, in which PHI was not purged from a photocopier before its return to the lessee, falls under both Level 2 and Level 3 errors. The company's employees not only lacked understanding of how information was stored in the photocopier, but they also failed to project the

consequences of the subsequent lease of the equipment, which resulted in exposure of PHI. Mental projection is a cognitively demanding task, and people generally perform poorly at it [13]. Examples of Level 3 SA errors are shown in Fig. 25.7.

By applying Endsley's [8] model to HIPAA privacy breach incidents, we are able to demonstrate the importance of improving the SA of employees through training or well-designed user interfaces so that such actors can correctly perceive, comprehend, and project a situation and minimize inadvertent or harmful disclosures of PHI. Although prior training, experience, and adequate monitoring contribute to reductions in SA errors, high workloads, understaffing, and multitasking may have negative effects and may result in disclosure of PHI.

25.4 Discussion and Conclusion

SA plays an important role in HIPAA compliance, especially as part of an error avoidance strategy that focuses on preventing human errors or at least on reducing their frequency. A mismatch between a worker's mental understanding of a system and the system's actual status is often linked to error. In such cases, poor feedback and lack of experience are considered the main causes of human error. To avoid such errors, frequent training of users to enhance both their knowledge and skills is highly appropriate. It is also desirable to have such recurrent training integrated into a system's normal operations. Similarly, user interface design has a significant role in avoidance of errors. Privacy-enhancing technologies (PET) have gained considerable prominence in recent times. Whenever information is handled electronically, an information space provides a way to organize information, resources, and services around important privacy-relevant contextual factors. Barcodes and other memory aids are also used to enhance awareness so as to avoid errors. All of these solutions are closely linked to SA enhancement. It is clear that achieving a high level of SA is crucial to the prevention of SA errors.

In summary, to reduce the possibility of SA errors, a number of best practices that apply to all three levels are suggested. First, under the category of awareness and education, employees who handle PHI and their supervisors should have easy access to HIPAA guidelines and be subject to frequent awareness campaigns as well as compliance training and evaluation [5]. In the category of administrative controls, covered entities must institute frequent audits of privacy practices and have supervisors periodically examine employees' work practices. They also should reduce work overloads, multitasking, and interruptions in workflow and take timely and exemplary action upon detection of violations. A third category encompasses technical controls that include increased workflow automation, privacy controls in application software that handles PHI, and enhanced SA aids such as alarms and decision support tools. Liginlal et al. [15, 16] discuss in greater detail these best practices for mitigating human errors that lead to privacy breaches.

Privacy management within an organization is only one aspect of overall privacy-related issues. Business transactions and housekeeping often require the exchange of private information belonging to both patients and employees. With the popularity

of highly networked business environments, individuals are frequently required to make quick decisions with significant implications for privacy [6, 18]. Given the highly interwoven and complicated environment the World Wide Web provides, it is not easy for organizations to provide individual users with detailed policies and timely instructions. When individuals fail to give proper attention to situational factors, they are likely to commit errors that lead to privacy breaches. Higher levels of multitasking, distraction, fatigue, or workloads experienced by users contribute to a lower level of SA, eventually culminating in inadvertent and often harmful disclosures of personal information. Therefore, determining whether a specific act of disclosing information adequately meets both the individual's need to share information and his or her privacy protection requirements is a complex task involving situated cognition.

There has always existed a concern that an individual's privacy protective behavior might inhibit the free flow of personal health information requisite to delivering quality care. With the increasing need for individuals to disclose their personal health information online or have their vital signs monitored by wearable or other mobile devices, a much neglected facet of an individual's privacy protective behavior also becomes important. Applying SA to such personal information disclosure to address a person's ability to deal with a situated environment under conditions of limited resources, such as a low attention span and limited short-term memory, seems reasonable and meaningful. Also, as social exchange and privacy regulation theories [1, 7] suggest, privacy needs and consequent privacy protective behaviors are specific to users and to the environment. A user's experience and reputation, goals and expectations, the degree of disclosure to peers within peer groups—with consequent pressure to reveal or hide—are all considered in those decisions related to the boundaries of privacy disclosure. When it comes to online disclosure of personal health information, it is often likely that users will not know precisely the audience for their information. For example, when a patient is using a chat feature on a healthcare website or using a mobile phone to converse with a healthcare call center, the surroundings of the patient influence the richness of information the patient is able to divulge. One also cannot rule out the likelihood that the operator at the other end is not using a headset and consequently, others may overhear the conversation. Thus, an understanding of the situation and awareness of the features of the system are both important factors that emphasize the central role of SA in such contexts. Thus, by introducing SA within the framework of the privacy calculus model, the often neglected situated aspects of individuals' privacy behaviors can be better understood and more easily predicted.

In conclusion, the instrumentation of SA applied to information privacy yields better insights into designing privacy-enhancing user interfaces, powerful help features, and privacy alerts. In addition to its applicability to healthcare settings, the SA construct can be easily applied to other contexts such as e-commerce and other organizational settings [4, 22]. Organizations and their employees are constrained to show more protective and conservative behavior. Organizational actions such as training, policies, and the improvement of environmental and business processes that are aimed at preventing human error are all important and highly recommended.

However, if the individuals who manage sensitive information do not perform as expected, the associated problems of human errors will not be solved. Therefore, attention to individuals and an understanding of how and why they commit errors are important facets of managing and preventing human errors that lead to privacy breaches in organizations.

References

1. Altman, I.: *The Environment and Social Behavior: Privacy, Personal Space, Territory, Crowding*. Brooks/Cole, Monterey (1975)
2. Annas, G.: HIPAA regulations – a new era of medical-record privacy? *N. Engl. J. Med.* **348**(15), 1486–1490 (2003)
3. Anton, A., Qingfeng, H., Baumer, D.: Inside JetBlue’s privacy policy violations. *IEEE Secur. Priv.* **2**(6), 12–18 (2004)
4. Berendt, B., Gunther, O., Spiekermann, S.: Privacy in e-Commerce: stated preferences vs. actual behavior. *Commun. ACM* **48**, 38–51 (2005)
5. Blumenthal, D., McGraw, D.: Keeping personal health information safe: the importance of good data hygiene. *J. Am. Med. Assoc.* **313**(14), 1424–1424 (2015)
6. Culnan, M., Armstrong, P.: Information privacy concerns, procedural fairness, and impersonal trust: an empirical investigation. *Organ. Sci.* **10**(1), 104–115 (1999)
7. Dinev, T., Hart, P.: An extended privacy calculus model for e-Commerce transactions. *Inf. Syst. Res.* **17**(1), 61–80 (2006)
8. Endsley, M.: Situation awareness in aviation systems. In: Garland, D.J., Wise, J.A., Hopkin, V.D. (eds.) *Handbook of Aviation Human Factors*. CRC Press, Boca Raton (1999)
9. Erickson, J., Millar, S.: Caring for patients while respecting their privacy: renewing our commitment. DOI: [10.3912/OJIN.Vol10No02Man04](https://doi.org/10.3912/OJIN.Vol10No02Man04) *Online J. Issues Nurs.* **10**(2) (2005)
10. HITECH Act: 42 USC 139w-4(0)(2). <http://www.hhs.gov/ocr/privacy/hipaa/administrative/enforcementrule/hitechenforcementifr.html> (2009). Accessed 27 Mar 2015
11. Ishikawa, K., Ohmichi, H., Umesato, Y., Terasaki, H., Tsukuma, H., Iwata, N., Tanaka, T., Kawamura, A., Sakata, K., Sainohara, T., Sugimura, M., Konishi, N., Umemoto, R., Mase, S., Takesue, S., Tooya, M.: The guideline of the personal health data structure to secure safety healthcare. The balance between use and protection to satisfy the patients’ needs. *Int. J. Med. Inform.* **76**(5), 412–418 (2007)
12. Johnson, M., Goetz, E.: Embedding information security into the organization. *IEEE Secur. Priv.* **5**(3), 16–24 (2007)
13. Jones, D., Endsley, M.: Sources of situation awareness errors in aviation. *Aviat. Space Environ. Med.* **67**(6), 507–512 (1996)
14. Kagal, L., Abelson, H.: Access control is an inadequate framework for privacy protection. In: *Proceedings of the W3C Workshop on Privacy for Advanced Web APIs*, Paper No. 20 (2010)
15. Liginlal, D., Sim, I., Khansa, L.: How significant is human error as a cause of privacy breaches? An empirical study and a framework for error management. *Comput. Secur.* **28**(3–4), 215–228 (2009)
16. Liginlal, D., Sim, I., Khansa, L., Fearn, P.: HIPAA privacy rule compliance: an interpretive study using Norman’s action theory. *Comput. Secur.* **31**(2), 206–220 (2012)
17. McCann, E.: Biggest health data breaches, Healthcare IT News. <http://www.healthcareitnews.com/slideshow/slideshow-top-10-biggest-hipaa-breaches> (2015). Accessed 20 April 2015
18. Otto, P., Anton, A., Baumer, D.: The ChoicePoint dilemma: how data brokers should handle the privacy of personal information. *IEEE Secur. Priv.* **5**(5), 15–23 (2007)
19. Ponemon Institute: IBM 2015 cost of data breach study - global analysis. <http://www-03.ibm.com/security/data-breach/> (2015). Accessed 22 June 2015
20. Reason, R.: *Human Error*. Cambridge University Press, New York (1990)

21. Sim, I.: Online Information Privacy and Privacy Protective Behavior: How Does Situation Awareness Matter? Ph.D. Dissertation, The University of Wisconsin-Madison (2010)
22. Sim, I., Liginlal, D., Khansa, L.: Information privacy situation awareness: construct and validation. *J. Comput. Inf. Syst.* **53**(1), 57 (2012)
23. Smith, C.: Somebody's watching me: Protecting patient privacy in de-identified prescription health information. *Vermont Law Rev.* **36**, 931 (2011)
24. US Department of Health & Human Services: Breaches affecting 500 or more individuals. https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf (2015). Accessed 30 April 2015
25. US Department of Health & Human Services: Case examples and resolution agreements. <http://www.hhs.gov/ocr/privacy/hipaa/enforcement/examples> (2015). Accessed 30 April 2015
26. US Department of Health & Human Services: Health information privacy. <http://www.hhs.gov/ocr/privacy/hipaa/understanding/consumers/index.html> (2015). Accessed 30 April 2015
27. Westin, A.: Privacy and Freedom. Bodley Head, New York (1967)

Chapter 26

De-identification of Unstructured Clinical Data for Patient Privacy Protection

Stephane M. Meystre

Abstract The adoption of Electronic Health Record (EHR) systems is growing at a fast pace in the United States and in Europe, and this growth results in very large quantities of patient clinical information becoming available in electronic format, with tremendous potential, but also equally growing concern for patient confidentiality breaches. Secondary use of clinical information is essential to fulfil the promises for high quality healthcare, improved healthcare management, and effective clinical research. De-identification of patient information has been proposed as a solution to both facilitate secondary use of clinical information, and protect patient information confidentiality. Most clinical information found in the EHR is unstructured and represented as narrative text, and de-identification of clinical text is a tedious and costly manual endeavor. Automated approaches based on Natural Language Processing have been implemented and evaluated, allowing for much faster de-identification than manual approaches. This chapter introduces clinical text-de-identification in general, and then focuses on recent efforts and studies at the U.S. Veterans Health Administration. It includes the origins and definition of text de-identification in the United States and Europe and a discussion about text anonymization. It also presents methods applied for text de-identification, examples of clinical text de-identification applications, and U.S. Veterans Health Administration clinical text de-identification efforts.

26.1 Introduction

Electronic Health Record (EHR) systems' adoption is growing at a fast pace in the United States and in Europe, and this evolution results in very large quantities of patient clinical information becoming available in electronic format, with remarkable potential, but also rising concern for patient confidentiality infringements. In the U.S., the adoption of EHR systems is growing fast, reaching more than 50 % of physician practices and 80 % of hospitals in April 2013, when only 17 % of

S.M. Meystre (✉)

Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA

e-mail: stephane.meystre@hsc.utah.edu

physician practices and 9 % of hospitals were using an EHR in 2008 [52]. This fast growth has been encouraged by the Health Information Technology for Economic and Clinical Health (HITECH) Act, which authorized incentive payments [7]. These payments are made through Medicare and Medicaid to clinicians and hospitals when using EHRs according to meaningful use requirements. Perspectives of improved healthcare quality also fuelled this fast growth. In Europe, most EHR adoption efforts have been encouraged by national health improvement efforts such as the Swedish National Patient Summary (NPO) initiative [48], the Danish Health Data Network [33] or the Australian Personally Controlled Electronic Health Record (PCEHR) initiative [58].

Secondary use of clinical information is essential to fulfil the promises for high quality healthcare, improved healthcare management, and effective clinical research. Instead of relying on existing but often biased and insufficiently detailed diagnostic and procedure codes assigned for reimbursement and administrative purposes only, effective clinical research and high quality and efficient healthcare need accurate and detailed clinical information, information that can be found in patient EHRs. Access to rich and detailed clinical information about diagnoses, treatments, and outcomes is also required for the Positive predictive value Medicine proposed by the National Academy of Sciences [40].

De-identification of patient information has been proposed as a solution to both facilitate secondary uses of clinical information, and protect patient information confidentiality. For flexibility, expressiveness, efficiency, and historical reasons, most detailed clinical information found in EHRs is captured in free-text format, without structure or coding. In a recent survey of U.S. hospitals equipped with advanced EHRs, only an average of 35 % of their clinical data was captured using structured templates, while the remaining used dictation and transcription or direct entry, both resulting in free text content [8]. De-identification of clinical text is a tedious and costly manual endeavor. Automated approaches based on Natural Language Processing (NLP) have been implemented and evaluated, allowing for much faster de-identification than manual approaches.

This chapter continues with a definition of text de-identification and a presentation of its origins. The main methods applied for text de-identification and three examples of de-identification systems are then discussed. The question of text anonymization is considered, and finally clinical text de-identification efforts at the U.S. Veterans Health Administration are presented.

26.2 Origins and Definition of Text De-identification

Confidentiality of the information entrusted by a patient to a healthcare provider has been a foundation of the confidence relationship established between them for centuries, as expressed in the Hippocratic Oath in ancient Greece: “... All that may come to my knowledge in the exercise of my profession or in daily commerce with

men, which ought not to be spread abroad, I will keep secret and will never reveal” [39]. Breaching this confidentiality not only damages this relationship, but also exposes the patient to financial, reputation, employment, and other identity theft disastrous consequences. The patient may become liable for services fraudulently obtained in his name, and incorrect information can infiltrate his medical record, potentially even threatening his life if his blood type was altered for example, or important medication allergies were erroneously removed. The patient may have damaged credit history, higher health insurance premiums, loss of health insurance coverage, and even legal troubles because of a stolen medical identify. The U.S. Internal Revenue Service (IRS) had identified almost 642,000 identity theft incidents that impacted tax administration in 2012 alone, up from 245,000 in 2010 [54].

In the United States, the 1996 Health Insurance Portability and Accountability Act (HIPAA) and 2000 Privacy Rule (codified as 45 CFR §160 and 164) protect the confidentiality of patient data [26] and require notification of breach of Protected Health Information (PHI) incidents. As of September 30 2013, the Department of Health and Human Services has received more than 86,000 privacy rule complaints [53] and lists 712 reports of incidents involving 500 individuals or more, including more than 27 million patient files and averaging an astonishing 19,000 patient files lost every day! [51]

In the European Union, the European Convention on Human Rights and the Data Protection Directive Article 8 offer similar legal bases, with corresponding national legislations in each member states. Directive 95/46/EC of the European Parliament defines “personal data” as “any information relating to an identified or identifiable natural person (‘data subject’); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity” [13].

The protection of data privacy is regulated in all European Union member states by national legislation drawn up in response to the aforementioned Data Protection Directive, which seeks to harmonize the rules of data protection throughout the European Union.

These laws typically require the informed consent of the patient and approval of a local ethics or internal review board to use data for research purposes, but these requirements could be waived if data are de-identified. In the U.S., for clinical data to be considered de-identified, the HIPAA “Safe Harbor” standard requires 18 data elements (called *PHI*: Protected Health Information) to be removed (see Fig. 26.1). In the European Union, no equivalent de-identification legislation is available. Anonymization has been defined within the law of some member states, like Germany, where it is stated that “the modification of personal data so that the information concerning personal or material circumstances can no longer or only with a disproportionate amount of time, expense and labor be attributed to an identified or identifiable individual” [16].

1. Names.
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, zip code or equivalents except for the initial 3 digits of a zip code if the corresponding zone contains more than 20, 000 people.
3. All elements of dates (except year) for dates directly related to the individual (birth date, admission date, discharge date, date of death). Also all ages over 89 or elements of dates indicating such an age.
4. Telephone numbers.
5. Fax numbers.
6. E-mail addresses.
7. Social security numbers.
8. Medical record numbers.
9. Health plan numbers.
10. Account numbers.
11. Certificate or license numbers.
12. Vehicle identification or serial numbers including license plate numbers.
13. Device identification or serial numbers.
14. universal resource locators (URLs).
15. Internet Protocol addresses (IP addresses).
16. Biometric identifiers.
17. Full face photographs and comparable images.
18. *Any other unique identifying number, characteristic, or code.*

Fig. 26.1 U.S. HIPAA protected health information categories

Directive 95/46/EC *does not apply to personal data rendered anonymous*. It states, “the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable” [13].

De-identification corresponds to the replacement or removal of personal identifiers so that linking an individual with his or her data becomes difficult. In the U.S., de-identification consists in the application of the HIPAA Safe Harbor standard, replacing or removing each mention of PHI (see Fig. 26.1).

De-identification of narrative text documents is often realized manually, and requires significant resources. Dorr et al. [14] have evaluated the time cost to manually de-identify narrative text notes (average of 87.2 seconds per note), and concluded that it was time-consuming and difficult to exclude all PHI required by HIPAA. Well aware of these issues, several authors have developed applications for automated de-identification of narrative text documents from the EHR using methods based on NLP, starting with Sweeny’s Scrub system in 1996 [49]. These applications focused on a variable selection of PHI, ranging from patient names only, to all PHI categories, or even everything that was not recognized as clinical information. Most applications focused on only one or two specific clinical document types, and only few systems were evaluated with a more heterogeneous collection of notes.

26.3 Methods Applied for Text De-identification

The text de-identification process is composed of two main steps: *PHI detection*, and *PHI removal or transformation* (see Fig. 26.2). The former is often approached as a “named entity recognition” task, and the latter consists of replacing PHI with some tags or characters (e.g., “Mr. Smith” becomes “<Patient_name>”). Another option is to replace PHI with synthetic but realistic substitutes (e.g., “Mr. Smith” becomes “Mr. Jones”). This second option is often called *PHI resynthesis* and has been experimented by Aberdeen and Yeniterzi et al. [1, 60] and within the U.S. Veterans Health Administration (VHA) clinical text de-identification project described below [20]. It adds computational complexity but offers the advantage of allowing the rare instances of PHI that could have been missed to “hide in plain sight,” notably improving the effectiveness of de- identification [9]. The remainder of this section will focus on PHI detection methods.

Automated text de-identification is mainly based on two different types of methodologies: pattern matching [4] and machine learning [37]. Characteristics of both types of methodologies are summarized in Table 26.1. Many systems combine both approaches for different types of PHI, but the majority uses no machine

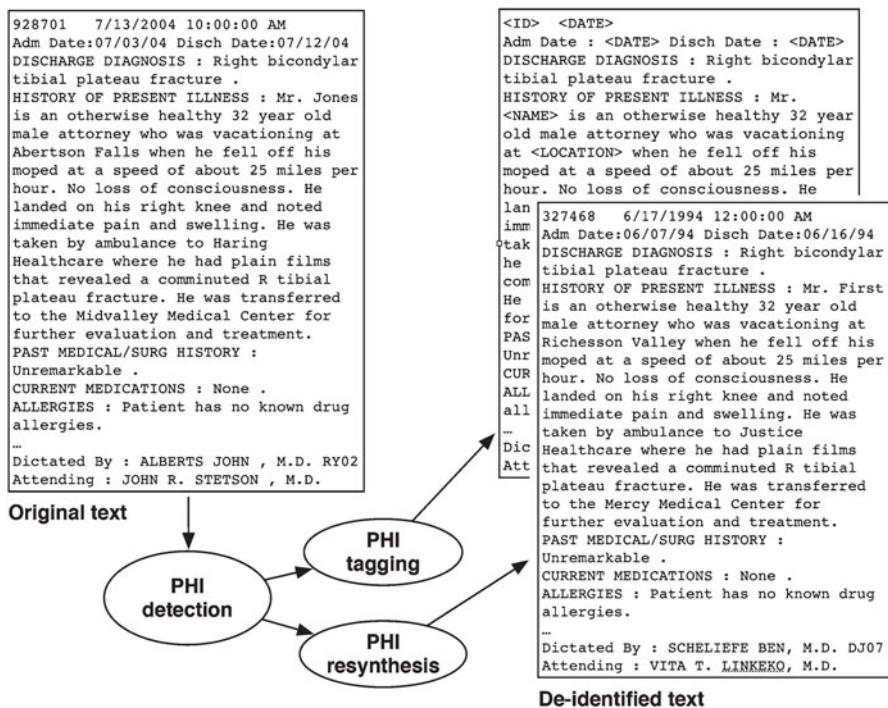


Fig. 26.2 Examples of text de-identification with PHI tagging or resynthesis

Table 26.1 Advantages and disadvantages of methods used for text de-identification

Characteristics	Pattern matching	Machine learning
Manually crafted knowledge	Yes	No
Manually annotated training data	Yes	Large quantity
Need to know all data variety	Yes	No
Small modifications	Easy	Difficult
Generalizability	Limited	Good
Common method examples	Regular expressions, dictionaries, rules	Support vector machines, conditional random fields, decision trees, Naïve Bayes
Types of PHI commonly detected	Dates, ages, phone numbers	Addresses, names

learning and relies on pattern matching, rules, and dictionaries. These resources are usually manually crafted by experienced domain experts, at high costs, and with limited generalizability.

Advantages of the rule-based and pattern matching de-identification methods are that they require little or no annotated training data, and that they can be easily modified to improve performance by adding rules, dictionary terms, or regular expressions. Pattern matching is often implemented with regular expressions (up to 50 or more for several systems) [23].

Dictionaries include terms typically considered PHI: proper names, locations, healthcare institution names, and sometimes even actual names of patients or healthcare providers in the institution the system was developed in [21, 24, 41]. These dictionaries are built from multiple sources such as the U.S. Social Security Death Index, spell-checking lexicons that include proper names (e.g., Ispell), or lists of geographical entities. For disambiguation, some systems also include dictionaries of terms that are in general not considered PHI (e.g., general English terms, biomedical terms, etc.).

Disadvantages of the rule-based and pattern matching methods are the aforementioned requirement for manually crafted multiple and complex algorithms to account for the different categories of PHI, and the requirement for developers to be aware of all possible PHI patterns that could occur (e.g., unexpected date formats, non-standard abbreviations). Rule-based and pattern matching methods also may not generalize to different datasets (i.e., data from a different institution or a different type of medical reports) and require manual customization.

Good examples of applications using mostly rule-based and pattern matching methods include HMS Scrubber, an application developed by Beckwith et al. for pathology reports in the Shared Pathology Informatics Network (SPIN) [5]. It uses 50 regular expressions to detect PHI and replace it with tags indicating its category. Another good example is the “deid” software package developed by Neamatullah et al. for Physionet [41]. It uses lexical look-up tables, regular expressions, and rules to detect PHI and also replace it with tags indicating its category. A last example is “De-ID,” an application developed at the University of Pittsburgh, evaluated

by Gupta [28], and commercialized by De-Id Data Corp. It uses rules, pattern matching algorithms, and dictionaries to identify PHI. These dictionaries include patient demographic information extracted from report headers. Detected PHI is then replaced with tags indicating its category.

Among recent text de-identification applications, the majority tends to be mostly based on supervised machine learning algorithms such as Support Vector Machines [47], Conditional Random Fields [32], Decision Trees [44], or Maximum Entropy [46]. These algorithms require a large collection of annotated text for training, a resource that also requires significant work by domain experts, although text annotation is often considered to be easier than knowledge engineering. Annotated collections of text can also be shared as organized during the i2b2 de-identification challenge [55]. Most systems based on machine learning also add some pattern matching to detect PHI that tends to be regular such as telephone numbers or social security numbers. For example, the MITRE Identification Scrubber Toolkit (presented in more details below [1]) uses regular expressions to capture telephone numbers, ZIP codes, dates, and other numeric identifiers, and then uses them as features for their conditional random fields algorithm. Another example is the system developed by Szarvas and colleagues [50] that uses regular expressions to capture age mentions, telephone numbers, dates, and other numeric identifiers, to then uses them with their decision tree algorithm. As shown in these examples, pattern matching is typically applied to extract features used by the machine learning algorithms.

The variety of features used with machine learning algorithms includes lexical, syntactic, semantic, document-level, and collection-level features. Lexical features (word-level features) describe the word case, punctuation, special characters, numerical characters, and the morphology of the word (e.g., “Aa” to indicate words starting with a capital letter). They are by far the most common type of features used for de-identification, and are also described as the most useful for the classification task. Syntactic features almost always include the part-of-speech of the words (e.g., “NN” to indicate a noun). Semantic features refer to the semantic classification of word or phrases and include terms from dictionaries, and semantic types (e.g., “C0004096” to indicate “Asthma” in the UMLS Metathesaurus). Document-level features are less common and include section headers or frequency of terms in the document (e.g., “FamHx” to indicate that the term was found in a “Family History” section). Collection-level features are rarely used and include word frequencies, mostly for disambiguation.

Advantages of machine learning de-identification methods include their ability to automatically learn to recognize complex PHI patterns, with developers of the system only requiring little knowledge of PHI patterns. When adapted to new document types or domains, machine learning-based de-identification systems tend to increase less in complexity and see their processing speed slow down less than pattern matching-based systems.

Their main disadvantage is the already mentioned need for large amounts of annotated training data (for supervised learning [29]). Although machine learning-based de-identification systems are typically more generalizable than pattern

matching-based systems, some additional annotated training is often required when applied to a new domain. Another disadvantage of machine learning is that it is sometimes difficult to know the reasoning and resources used when the application erroneously detects or misses PHI. This “black box” problem is reduced when using algorithms like decision trees or rule learners.

Good examples of applications using mainly machine learning methods include the Health Information DE-identification (HIDE) system developed by Gardner et al. [25]. It uses Conditional Random Fields in an iterative classifying and retagging process during development of the system and was focused on pathology reports. Another good example is the MITRE Identification Scrubber Toolkit (MIST) developed by Aberdeen et al. [1]. MIST is based on a tool implementing Conditional Random Fields for text processing—Carafe [46]—and is described in more details below. Finally, “BoB”, a Best-of-Breed clinical text de-identification system for the VHA is also a good example. It was developed by Fernandez, et al. [20] and combines machine learning with pattern matching and rules in a stepwise approach explained in more details below.

In general, methods based on dictionaries perform better with PHI that is rarely mentioned in clinical text, but are more difficult to generalize [19]. Methods based on machine learning tend to perform better overall, especially with PHI that is not mentioned in the dictionaries used [19]. In the i2b2 de-identification challenge [55, 56], systems based on machine learning with regular expression template features for all PHI categories performed best. They were followed by systems combining rules for some PHI categories with learning for others, then by systems purely based on machine learning without regular expression template features or rules, and finally by purely rule-based systems [55].

For a more detailed analysis of methods used for text de-identification, please refer to publications by Meystre and colleagues [34] and by Kushida [31], with applications to unstructured data other than text like images and biosamples.

26.4 Clinical Text De-identification Application Examples

Three examples of text de-identification applications are presented here, selected for their originality, performance, or availability: Physionet’s deid [41], MITRE’s MIST [1], and the VHA best-of-breed (BoB) clinical text de-identification system [20]. This presentation includes details of the resources and methods used, as well as performance evaluations.

26.4.1 *Physionet Deid*

Physionet deid was developed and evaluated with a large collection of clinical notes from the MIMIC II database, a large database of cardio-vascular and related signals

and accompanying clinical data from intensive care units in the United States. The deid application is based on lexical look-up tables, regular expressions, and simple rules to identify PHI in medical text documents. It is developed in Perl and uses four types of look-up dictionaries: names of patients and hospital staff (obtained from the MIMIC II database), PHI names and locations which are also medical or common words (i.e., words that are in the UMLS Metathesaurus or a spell-checking dictionary, keywords and phrases that likely act as indicators for PHI (“Mr.”, “Dr.”, “hospital”, “Street”, etc.), and common words and phrases likely to be non-PHI (obtained from the UMLS Metathesaurus and the spell-checking dictionary). Numerical PHI patterns (e.g., phone numbers) are identified with regular expressions, and non-numerical PHI (e.g., patient names) is identified with a combination of dictionaries, a context detection algorithm, and regular expressions.

The Deid system was evaluated with 2434 clinical notes from 163 patients randomly selected from the MIMIC II database [41]. The notes contained 1779 instances of PHI as identified by the reference standard of three clinician reviewers and a fourth to adjudicate disagreements. The system missed 59 instances of PHI, with an estimated sensitivity of 96.7 % (average of all PHI types). Missed PHI was mostly dates and location information. The system did commit numerous false positive errors resulting in a positive predictive value of only 75 % (average of all PHI types). However, it was noted that the readability and information content of the de-identified notes was not compromised despite the low positive predictive value.

Deid is available with an open source license (GNU Public License version 2), and several authors adapted it to their own needs. Grouin et al. combined it with resources inspired from MeDS [24] to de-identify French clinical text [27]. This Deid adaptation then inspired development of a new system, Medina, also using regular expressions and dictionaries. The latter were adapted to French. When evaluated with 23 randomly selected notes from a corpus of 21, 749 clinical notes, these systems reached good accuracy (91 % sensitivity, 100 % positive predictive value) with de-identification and removal of patient name and birth date, and with Medina (83 %, 92 %), and lower accuracy with the French adaptation of Deid (65 %, 23 %). Morrison combined it with MedLEE to improve de-identification accuracy [38]. Adding MedLEE after Deid allowed missing only 2.1 % of PHI instead of 24 % when using Deid alone. Vellupilai and Dalianis adapted Deid to Swedish clinical text [57]. Deid-Swe was adapted to detect the Swedish phone number and social security number formats, and lists of Swedish terms were added. When evaluated with a corpus of 100 clinical notes, it reached an F1-measure of 80 % for names, but produced many false positives for other categories of PHI; its positive predictive value was only 3–9 % and the sensitivity was 56–76 %.

26.4.2 MIST (*MITRE Identification Scrubber Toolkit*)

MIST, the MITRE Identification Scrubber Toolkit, was developed as a complete system for rapid customization to new clinical text domains. It includes a web-based

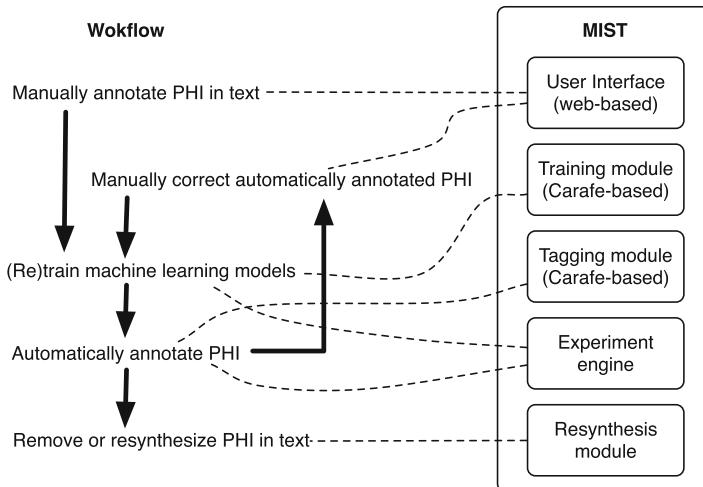


Fig. 26.3 Workflow supported by MIST and its components

graphical user interface for PHI annotation in clinical text. Other components are a training module, a tagging module, a re-synthesis module, and an experiment engine (see Fig. 26.3). The training module is based on Conditional Random Fields (Carafe implementation) [59] and uses PHI annotations to learn models for each type of PHI. The tagging module uses the trained models to detect PHI in the text. The resynthesis module then replaces the detected PHI with realistic surrogates. Finally, the experiment engine allows iterative training-testing cycles to adapt the system to a new clinical text domain. MIST is described in details in [1].

When trained and tested with 1200 clinical notes of various types from the Vanderbilt University Medical Center, MIST reached an overall sensitivity of 97.8 % and positive predictive value of 94.3 % [1].

MIST is available with open source licenses. It was used in the U.K. by Fernandes, Callard, and colleagues to partly de-identify mental health clinical notes. When trained with 50 clinical notes to de-identify first and last names only, and evaluated with 20 clinical notes, 78.1 % sensitivity and 95.1 % positive predictive value were measured. The authors compared it with a simple dictionary-based system that matched existing structured PHI found in the EHR, a system that reached 88.5 % sensitivity and 100 % positive predictive value [17].

26.4.3 VHA Best-of-Breed Clinical Text De-identification System

The VHA best-of-breed clinical text de-identification system (BoB) was developed and evaluated in the study described in the next section. As indicated in its name, BoB combines best-of-breed methods and resources for each type of PHI.

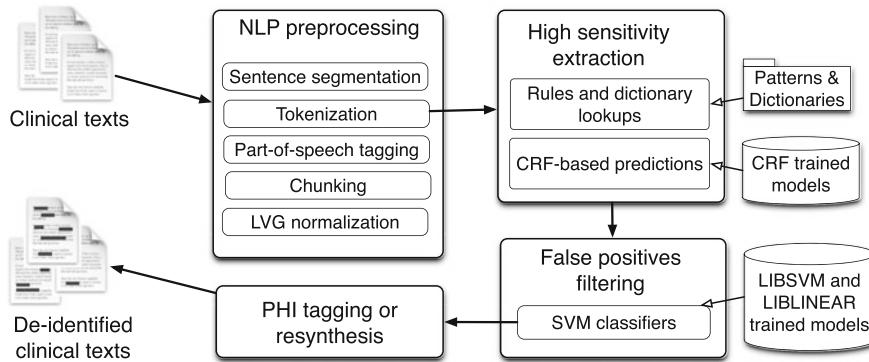


Fig. 26.4 The VHA BoB components and text processing pipeline

Knowledge of the methods and resources performing the best was based on a large literature and technology review [34], and evaluations of several existing text de-identification applications on various clinical text corpora (more details in the last section) [18, 19]. These evaluations demonstrated that no existing system reached sufficient accuracy when de-identifying VHA clinical text, even when trained with such clinical text. These findings motivated the development and evaluation of a new system for VHA clinical text de-identification, a system combining methods and resources allowing for the best detection accuracy with each type of PHI.

High sensitivity and high positive predictive value are often not compatible when detecting or extracting information from text. A highly sensitive system typically has lower positive predictive value (i.e., produces more false positives), and a highly precise system usually has lower sensitivity (i.e., more false negatives). Text de-identification requires very high sensitivity, but too many false positives could damage the non-PHI information in clinical notes. To try and combine very high sensitivity with high positive value, we developed a “hybrid stepwise” approach for BoB. After pre-processing of the clinical text, this approach consists in starting with a component focused only on high sensitivity, even if producing numerous false positives, and then continuing with a filtering component that filters out false positives (please see Fig. 26.4).

Pre-processing includes segmentation of the text in sentences, then followed by tokenization, part-of-speech tagging, noun phrase chunking, and term normalization. These functionalities were adapted from OpenNLP [43], with trained models and the implementation of LVG found in Apache cTAKES [2].

The high sensitivity extraction component combines various methods to detect all possible PHI in clinical text. A collection of about 130 regular expressions were adapted from two existing de-identification systems [24, 41], and enriched with new expressions to cover different PHI formats found in VHA documents (e.g., date-time formats such as “09/09/09@1200”). Dictionaries were used with Apache Lucene [3], implementing exact and fuzzy searches on dictionaries of first and last names

(from the 1990 U.S. census), U.S. states, cities, and counties, countries, companies (from Wikipedia, usps.com, and other web resources), common words (from Neamatullah et al. [41]), clinical eponyms, and healthcare clinic names extracted from VHA clinical notes. A machine learning module based on Conditional Random Fields (CRF, from the Stanford NLP group [22]) follows, aiming at detecting PHI missed by pattern matching and not found in dictionaries. This module includes classifiers for person names, addresses, healthcare and other organizations, and dates.

The “candidate PHI” produced by the high sensitivity extraction component is then passed to the false positives filtering component. The latter is composed of various Support Vector Machine (SVM, from LIBSVM [11]) classifiers. These classifiers included a multi-class SVM and four binary SVMs, all trained with the “candidate PHI” annotations from the high-sensitivity extraction component and focused on categories of PHI with more false positives: person names, numerical identifiers (e.g., dates, phone numbers, other ID numbers), and clinical eponyms. Clinical eponyms (i.e., clinical information that bears person names like Alzheimer’s disease or Achilles tendon) are not PHI, but could easily be misclassified as person names and deserve specific disambiguation.

The main evaluation of BoB was based on a corpus of 800 VHA clinical notes annotated for PHI by experts, and split in sub-corpora of 500 notes for training, and 300 notes for testing. Measured accuracy reached an average sensitivity of 92.6 % (98–100 % for person names, social security numbers, e-mail addresses, and ZIP codes), and 84.1 % positive predictive value [20].

26.5 Why Not Anonymize Clinical Text?

As defined above, de-identification means removing or replacing personal identifiers to make it difficult to reestablish a link between the individual and his or her data, but it doesn’t make this link impossible. Scrubbing and sanitization are sometimes used as synonyms for de-identification. In contrast, anonymization corresponds to the irreversible removal of the link between an individual and his or her data. Anonymization is sometimes used interchangeably with de-identification, but the former renders identification of the individual impossible, when the latter only makes this identification more difficult.

Even if perfectly de-identified, clinical narrative text is still rich in clinical information (as it should be). This information includes the patient family and social history and specific clinical information that can be unique and make identification of the patient a possibility. This risk is potentially significant. When including only the patient age in years, the gender, the first three digits of the ZIP (postal) code, the time in days since the last visit, and the length of stay at the hospital, 18.5 % of the approximately 1.5 million patient records in the New York state inpatient database for 2007 were unique [12]. At another extreme, a study based on voter registration information (i.e., date of birth, year of birth, race, gender, and county

of residence) estimated the re-identification risk to be between 0 % and 0.25 % in specific U.S. states, with wide variations depending on the information available [6]. Uniqueness has been one of the methods used to estimate the risk for structured and coded data (e.g., diagnostic codes, demographic information) re-identification, for linking a patient identity with his or her de-identified structured and coded data. It indicates the maximal theoretical risk, but this risk is mitigated by access difficulties to other data sets that include this clinical information along with the patient identity. Methods used to assess the risk for re-identification were applied to a small number of structured and coded data only, not to narrative text [15]. Considering the richness and variability of unique information that can be found in de-identified clinical notes, applying the same methods would require tremendous computing power, if not reveal intractable.

A realistic estimation of the risk for re-identification of automatically de-identified clinical notes has been realized in a U.S. Veterans Affairs clinical text de-identification project described in the next section [36]. It measured a very low risk for re-identification.

The anonymization of clinical text would require the removal or transformation of a large part of the patient clinical and social information, to ensure the absence of unique patient records in a given patient population. Only limited research has been realized in this field. The ERASE text sanitization system features a K-safety approach, modeling a document as a set of terms, and applying various algorithms to detect and then remove protected terms. An evaluation based on randomly-generated “documents” of 100 terms assessed the impact of various factors on processing time and proportion of terms retained, but no evaluation on realistic text documents was realized [10].

Another approach, t-plausibility, proposes heuristic algorithms to replace protected information found in text documents with more general information (e.g., “Denver” replaced with “state capital”, “marijuana” replaced with “drug”) instead of removing or hiding it. In the experimental evaluation, the authors used only 50 randomly selected words to assess the effect of their algorithms on processing time and similarity of sanitized words sets [30].

In general, if applying these approaches to reach anonymity, the removal or transformation of information would cause an important loss in clinical information, making the anonymized text almost or completely unusable for any subsequent usage, such as clinical research. Considering the very limited risk for re-identification of de-identified clinical text, anonymizing it would only offer limited interest and cause serious damage.

26.6 U.S. Veterans Health Administration Clinical Text De-identification Efforts

The U.S. Veterans Healthcare Administration (VHA) Consortium for Healthcare Informatics Research (CHIR) is a multi-disciplinary group of collaborating investigators affiliated with VHA sites across the U.S. In the context of the CHIR,

a de-identification project focused on investigating the current state of the art of automatic clinical text de-identification [19, 34], on developing a best-of-breed de-identification application for VHA clinical documents (presented above [20]) and on evaluating its impact on subsequent text analysis tasks [35] and the risk for re-identification of this de-identified text.

The state of the art of automatic clinical text de-identification investigation included an extensive literature review, analysis of available software systems architecture, and testing of these systems. The literature and technology search returned more than 200 scientific publications, but only 18 studies described automatic text de-identification and were included in our study. We observed that many existing systems focused on only a selection of PHI types defined in the HIPAA Safe Harbor standard, and most systems used only one, or few, clinical note types. Methods that performed well included machine learning approaches based on CRF, Decision Trees, Maximum Entropy models, or SVM, combined with dictionaries and sometimes regular expressions. This study produced rich information about the methods and resources allowing for best accuracy with each type of PHI. It also allowed selecting existing and available text de-identification applications for testing and comparative evaluation.

Three systems, mostly based on pattern matching and dictionaries (HMS scrubber [5], the Medical De-identification System (MeDS) [24], and the MIT deid software package [41]), and two machine learning-based systems (MITRE Identification Scrubber Toolkit (MIST) [1], and the Health Information DE-identification (HIDE) [25]) were installed locally and tested with the 2006 i2b2 de-identification challenge corpus [55], and a corpus of various VHA clinical documents [19]. When evaluated “out-of-the-box” with the VHA corpus, the pattern matching and dictionary-based systems only allowed for low accuracy, even if considering “fully-contained” partial matches (i.e., PHI found by the system at least overlapped with the entire PHI annotation in the reference standard). The sensitivity (see Fig. 26.5) of these systems averaged across all PHI types was measured between 0 % and 34 %, and the positive predictive value (see Fig. 26.5) between 0 % and 40 % only, but accuracy (see Fig. 26.5) was excellent with some PHI types like social security numbers and ZIP codes. Also, when considering partial matches, person names and other “textual” PHI were well de-identified by some rule-based systems (MeDS and the MIT deid system). The two machine learning-based systems were evaluated using a ten-fold cross-validation experiment, and obtained better positive predictive value in general. Their sensitivity varied between 0 % (for ages above 89) and 100 % (for ZIP codes), and increased when using dictionaries (Table 26.2). BoB reached the highest sensitivity in almost all categories (MIST had the highest sensitivity with healthcare unit names), demonstrating the need for capturing PHI features specific to VHA clinical text.

When comparing different configurations and uses of dictionaries at the PHI level (i.e., PHI or non-PHI, without considering the accuracy for each PHI type separately like in Table 26.1), BoB reached the highest sensitivity.

	Positive in reference standard	Negative in reference standard
Positive in output	True positives (TP)	False positives (FP)
Negative in output	False negatives (FN)	True Negatives (TN)

Sensitivity (S) = $\frac{TP}{TP+FN}$ Positive predictive value (PPV) = $\frac{TP}{TP+FP}$

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ F_β -measure [45] = $\frac{(1+\beta^2) \times S \times PPV}{(\beta^2 \times PPV) + S}$

$\beta = 1$ gives equal weight to S and PPV ; $\beta > 1$ gives more weight to S

Fig. 26.5 Typical metrics for de-identification applications

Table 26.2 Sensitivity of machine learning-based text de-identification applications (partly reproduced from [19])

PHI categories	MIST 1	HIDE 1	MIST 2	HIDE 2	MIST 3	BoB
Patient name	0.835	0.776	0.862	0.855	0.850	0.980
Relative name	0.440	0.360	0.520	0.840	0.440	0.920
Healthcare provider name	0.858	0.845	0.900	0.886	0.862	0.943
Other person name	0.333	0.555	0.555	0.555	0.555	0.888
Street or city	0.796	0.732	0.809	0.841	0.777	0.943
State or country	0.804	0.750	0.817	0.838	0.797	0.878
Deployment	0.868	0.811	0.849	0.773	0.773	0.887
ZIP code	1	1	1	0.800	1	1
Healthcare unit name	0.792	0.733	0.827	0.741	0.786	0.811
Other organization name	0.582	0.516	0.604	0.450	0.527	0.725
Date	0.960	0.940	0.963	0.934	0.956	0.971
Age above 89	0.250	0	0.250	0	0	1
Phone number	0.912	0.846	0.934	0.846	0.901	0.956
Electronic address	0.500	0.500	0	0.500	0	1
SSN	0.963	0.963	0.963	0.963	0.963	1
Other ID number	0.906	0.851	0.894	0.845	0.912	0.917
Overall (macro-avg) sensitivity	0.737	0.699	0.734	0.729	0.694	0.926

No dictionary: MIST 1 and HIDE 1; selection of dictionaries: MIST 2 and HIDE 2; all dictionaries: MIST 3 and BoB (not possible with HIDE)

HIDE (without dictionaries) reached the highest positive predictive value (Table 26.3). When giving more weight to sensitivity than to positive predictive value, BoB reached the highest F2-measure (0.904) (see Fig. 26.5).

To evaluate the generalizability of these machine learning-based systems, we trained them with the 2006 i2b2 de-identification challenge corpus, and tested them with the VHA clinical notes corpus [18]. This study demonstrated the limited

Table 26.3 Overall (micro-averaged) performance at the PHI level (partly reproduced from [19])

	MIST 1	HIDE 1	MIST 2	HIDE 2	MIST 3	BoB
Positive predictive value	0.926	0.936	0.715	0.933	0.700	0.836
Sensitivity	0.888	0.853	0.904	0.863	0.883	0.922
F_1 -measure	0.907	0.893	0.799	0.897	0.781	0.877
F_2 -measure	0.895	0.869	0.858	0.877	0.839	0.904

No dictionary: MIST 1 and HIDE 1; selection of dictionaries: MIST 2 and HIDE 2; all dictionaries: MIST 3 and BoB (not possible with HIDE)

Table 26.4 Sensitivity of de-identification applications when generalized to a different type of clinical notes

i2b2 PHI categories	MIST	HIDE	BoB
Patient	0.584	0.437	0.604
Doctor	0.549	0.392	0.600
Location	0.428	0.344	0.639
Hospital	0.645	0.487	0.839
Date	0.869	0.869	0.988
Age above 89	0	0	0
Phone number	0.948	0.810	0.845
ID	0.854	0.348	0.795
Overall macro-average sensitivity	0.610	0.461	0.664

generalizability of these text de-identification systems, even if some PHI types allowed for good accuracy (mostly numeric identifiers and dates; see Table 26.4).

When evaluating BoB, the problem of false positive errors altering clinical information became evident and motivated a study of the impact of de-identification on the informativeness and clinical information content of de-identified clinical notes [35]. After de-identification, most key clinical data and the overall meaning and understanding of documents were retained, even if one system removed all line spacing, paragraph breaks and indentation from the report, making it more difficult for the human reader to interpret and understand. To examine the impact on clinical information, counts of SNOMED-CT [42] concepts found by an open source information extraction application before and after de-identification of a corpus of VHA clinical notes were compared. The impact of de-identification was very limited, with only about 1.18–3.04 % less SNOMED-CT concepts found in de-identified versions of the VHA corpus. Many of these concepts were PHI erroneously identified as clinical information.

To study this impact in more details, we examined the overlap between problems, tests, and treatments annotated in the 2010 i2b2 NLP challenge corpus [56], and automatic PHI annotations from BoB. Overall, only 0.81 % of the clinical information exactly overlapped with PHI, and 1.78 % partly overlapped. Automated text de-identification's impact on clinical information is therefore small, but not negligible, and improved clinical acronyms and eponyms disambiguation could significantly reduce this impact.

Assessing the risk for de-identified text re-identification (i.e., establishing the identity of the patient discussed in the text) prove to be a difficult undertaking. Methods used to assess the risk for re-identification of de-identified data were applied to a small number of structured and coded data only (e.g., demographics, location) [6, 15], not to narrative text, and clinical documents are rich in clinical and social information that can be unique and could be used to re-identify a patient. Most automatic clinical text de-identification systems evaluated in the aforementioned studies allowed for a sensitivity of 88–99 % when detecting PHI [19, 34]. This means that some PHI is missed by these systems. Even if this is rare, would researchers using the de-identified clinical documents, or even healthcare providers who took care of the patients mentioned in the documents, recognize these patients when examining the documents?

To answer this question, a collection of 86 discharge summaries from patients hospitalized at the Salt Lake City VHA Medical Center acute medicine department (internal medicine and medical intensive care unit) between 1 and 3 months before the beginning of our study were automatically de-identified with PHI “resynthesis”. Physicians working in this department (five attending physicians, two chief medical residents, one subspecialty fellow, 11 medical residents) were then asked to read these de-identified notes and try identifying the patients described in the notes. Each note was examined by two physicians independently. Three physicians thought they had recognized a patient, citing reasons like specific procedures, diagnoses, signs, or imaging results. They often couldn't give any identifying information (e.g., patient name), and when they could, the information was compared with the correct patient identity. In the end, no “recognized” identity was correct and therefore no de-identified discharge summary could be re-identified.

As a conclusion, text de-identification methods and resources customized for VHA clinical text allowed for high accuracy detection and removal or transformation of PHI, with very limited impact on clinical notes readability, interpretability, and information content. The remaining clinical and social information, as well as PHI that could have been missed, didn't allow re-identifying any automatically de-identified discharge summaries, even when examined by the people most likely to recognize the patients described in these notes: physicians who worked in the hospital unit the patients were hospitalized in.

26.7 Conclusion

This chapter offered an overview of clinical text de-identification, from its definition and history, to methods used and examples of de-identification systems. Advantages and issues with pattern matching and rule-based methods versus machine learning-based approaches were discussed, and the clinical text de-identification studies realized at the U.S. Veterans Health Administration were explained in details.

For future efforts, limitations of text de-identification such as generalizability difficulties and lower accuracy with specific types of PHI (e.g., healthcare organization

names) offer opportunities for further research. The recognition of the theoretical risk for re-identification caused by remaining unique clinical information is a reason often cited for refusing re-release of de-identified data, even if early evidence indicates that this risk is very low, if not absent. Stronger evidence in this field would support increased de-identified clinical data sharing and availability, allowing for enhanced research based on this invaluable resource.

References

1. Aberdeen, J., Bayer, S., Yeniterzi, R., Wellner, B., Clark, C., Hanauer, D., Malin, B., Hirschman, L.: The MITRE identification scrubber toolkit: Design, training, and assessment. *Int. J. Med. Inform.* **79**(12), 849–859 (2010)
2. Apache cTAKES. <https://ctakes.apache.org> (2015). Accessed 20 June 2015
3. Apache Lucene. <http://lucene.apache.org/> (2015). Accessed 20 June 2015
4. Apostolico, A., Galil, Z.: Pattern Matching Algorithms. Oxford University Press, Oxford (1997)
5. Beckwith, B., Mahaadevan, R., Balis, U., Kuo, F.: Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC Med. Inform. Decis. Mak.* **6**, 12 (2006)
6. Benitez, K., Malin, B.: Evaluating re-identification risks with respect to the HIPAA privacy rule. *J. Am. Med. Inform. Assoc.* **17**(2), 169–177 (2010)
7. Blumenthal, D., Tavenner, M.: The “meaningful” use regulation for electronic health records. *N. Engl. J. Med.* **363**(6), 501–504 (2010)
8. Cannon, J., Lucci, S.: Transcription and EHRs. Benefits of a blended approach. *J. Am. Health Inf. Manag. Assoc.* **81**(2), 36–40 (2010)
9. Carrell, D., Malin, B., Aberdeen, J., Bayer, S., Clark, C., Wellner, B.: Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *J. Am. Med. Inform. Assoc.* **20**, 342–348 (2013)
10. Chakaravarthy, V., Gupta, H., Roy, P., Mohania, M.: Efficient techniques for document sanitization. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 843–852. ACM, New York (2008)
11. Chang, C.C., Lin, C.J.: LIBSVM : a library for support vector machines. <http://www.csie.ntu.edu.tw/~cjlin/libsvm> (2015). Accessed 20 June 2015
12. Dankar, F., El-Emam, K., Neisa, A., Roffey, T.: Estimating the re-identification risk of clinical data sets. *BMC Med. Inform. Decis. Mak.* **12**(1), 66 (2012)
13. Directive 95/46/EC of the European Parliament and of the Council: Eur-lex. 1995. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML> (1995). Accessed 24 July 2014
14. Dorr, D., Phillips, W., Phansalkar, S., Sims, S., Hurdle, J.: Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods Inf. Med.* **45**(3), 246–252 (2006)
15. El-Emam, K., Jonker, E., Arbuckle, L., Malin, B.: A systematic review of re-identification attacks on health data. *PLoS ONE* **6**(12), e28071 (2011)
16. Federal Data Protection Act. <http://www.iuscomp.org/gla/statutes/BDSG.htm> (2015). Accessed 20 June 2015
17. Fernandes, A., Cloete, D., Broadbent, M., Hayes, R., Chang, C.K., Jackson, R., Roberts, A., Tsang, J., Soncul, M., Liebscher, J., Stewart, R., Callard, F.: Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. *BMC Med. Inf. Decis. Mak.* **13**(1), 71 (2013)

18. Fernandez, O., South, B., Shen, S., Friedlin, F., Samore, M., Matthew, H., Meystre, S.: Generalizability and comparison of automatic clinical text de-identification methods and resources. *AMIA Annu. Symp. Proc.* **2012**, 199–208 (2012)
19. Fernandez, O., South, B., Shen, S., Friedlin, F., Samore, M., Meystre, S.: Evaluating current automatic de-identification methods with Veteran’s health administration clinical documents. *BMC Med. Res. Methodol.* **12**(1), 109 (2012)
20. Fernandez, O., South, B., Shen, S., Friedlin, F., Samore, M., Meystre, S.: BoB, a best-of-breed automated text de-identification system for VHA clinical documents. *J. Am. Med. Inform. Assoc.* **20**, 77–83 (2013)
21. Fielstein, E., Brown, S., Speroff, T.: Algorithmic de-identification of VA medical exam text for HIPAA privacy compliance: Preliminary findings. In: Proceedings of the 11th World Congress on Medical Informatics, p. 1590. Ios Press, Fairfax (2004)
22. Finkel, J., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 363–370. Association for Computational Linguistics, Stroudsburg (2005)
23. Friedl, J.: Mastering Regular Expressions. O’Reilly, Cambridge (2002)
24. Friedlin, F., McDonald, C.: A software tool for removing patient identifying information from clinical documents. *J. Am. Med. Inform. Assoc.* **15**, 601–610 (2008)
25. Gardner, J., Xiong, L., Li, K., Lu, J.: HIDE: Heterogeneous Information De-identification (2009). ACM, New York
26. GPO US. 45 C.F.R. S164: Security and privacy. http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr164_08.html (2008). Accessed 20 June 2015
27. Grouin, C., Rosier, A., Dameron, O., Zweigenbaum, P.: Testing tactics to localize de-identification. *Stud. Health Technol. Inform.* **150**, 735–739 (2009)
28. Gupta, D., Saul, M., Gilbertson, J.: Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. *Am. J. Clin. Pathol.* **121**(2), 176–186 (2004)
29. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York (2009)
30. Jiang, W., Murugesan, M., Clifton, C., Luo, S.: t-plausibility: semantic preserving text sanitization. In: Proceedings of the 2009 International Conference on Computational Science and Engineering, vol. 3, pp. 68–75 (2009)
31. Kushida, C., Nichols, D., Jadrnicek, R., Miller, R., Walsh, J., Griffin, K.: Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med. Care* **50**, 82–101 (2012)
32. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289. Morgan Kaufmann, San Francisco (2001)
33. MedCom: In english. <http://www.medcom.dk/wm109991> (2015). Accessed 20 June 2015
34. Meystre, S., Friedlin, F., South, B., Shen, S., Samore, M.: Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med. Res. Methodol.* **10**, 70 (2010)
35. Meystre, S., Fernandez, O., Friedlin, F., South, B., Shen, S., Samore, M.: Text de-identification for privacy protection: a study of its impact on clinical text information content. *J. Biomed. Inform.* **50**, 142–150 (2014)
36. Meystre, S., Shen, S., Hofmann, D., Gundlapalli, A.: Can physicians recognize their own patients in de-identified notes? *Stud. Health Technol. Inform.* **205**, 778–782 (2014)
37. Michell, T.: Machine Learning. McGraw-Hill, Maidenhead (1997)
38. Morrison, F., Sengupta, S., Hripcsak, G.: Using a pipeline to improve de-identification performance. *AMIA Annu. Symp. Proc.* **2009**, 447–451 (2009)
39. National Library of Medicine: The hippocratic oath. http://www.nlm.nih.gov/hmd/greek_greek_oath.html (2002). Accessed 20 June 2015

40. National Research Council (US): Committee on a framework for developing a new taxonomy of disease. In: *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. National Academies Press, Washington (2011)
41. Neamatullah, I., Douglass, M., Lehman, L., Reisner, A., Villarroel, M., Long, W., Szolovits, P., Moody, G., Mark, R., Clifford, G.: Automated de-identification of free-text medical records. *BMC Med. Inform. Decis. Mak.* **8**(1), 32 (2008)
42. NLM US. SNOMED Clinical Terms: SNOMED-CT. http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html (2015). Accessed 20 June 2015
43. OpenNLP. <http://opennlp.sourceforge.net/> (2015). Accessed 20 June 2015
44. Quinlan, J.: Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986)
45. Rijksenbergen, C.: *Information Retrieval*, 2nd edn. Butterworth-Heinemann, Newton (1979)
46. Rosenfeld, R.: A maximum entropy approach to adaptive statistical language modelling. *Comput. Speech Lang.* **10**(3), 187–228 (1996)
47. Steinwart, I., Christmann, A.: *Support Vector Machines*. Springer, New York (2008)
48. Sweden Jumps on National EHR Bandwagon: Healthitnewsdirect. <http://www.healthitnewsdirect.com/?p=116> (2009). Accessed 20 June 2015
49. Sweeney, L.: Replacing personally-identifying information in medical records, the Scrub system. In: *Proceedings: A conference of the American Medical Informatics Association*, pp. 333–337 (1996)
50. Szarvas, G., Farkas, R., Busa-Fekete, R.: State-of-the-art anonymization of medical records using an iterative machine learning framework. *J. Am. Med. Inform. Assoc.* **14**(5), 574–580 (2007)
51. U.S. Department of Health and Human Services: Breaches affecting 500 or more individuals. <http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrule/breatchool.html> (2015). Accessed 20 June 2015
52. U.S. Department of Health and Human Services: Doctors and hospitals' use of health IT more than doubles since 2012. <http://www.hhs.gov/news/press/2013pres/05/20130522a.html> (2015). Accessed 20 June 2015
53. U.S. Department of Health and Human Services: Numbers at a glance. <http://www.hhs.gov/ocr/privacy/hipaa/enforcement/highlights/indexnumbers.html> (2015). Accessed 20 June 2015
54. U.S. Government Accountability Office: Identity theft. <http://www.gao.gov/assets/660/650366.pdf> (2012). Accessed 20 June 2015
55. Uzuner, O., Luo, Y., Szolovits, P.: Evaluating the state-of-the-art in automatic de-identification. *J. Am. Med. Inform. Assoc.* **14**(5), 550–563 (2007)
56. Uzuner, O., South, B., Shen, S., DuVall, S.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J. Am. Med. Inform. Assoc.* **18**(5), 552–556 (2011)
57. Velupillai, S., Dalianis, H., Hassel, M., Nilsson, G.: Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. *Int. J. Med. Inform.* **78**(12), 19–26 (2009)
58. Welcome to eHealth.gov.au. <http://www.ehealth.gov.au/internet/ehealth/publishing.nsf/content/home> (2015). Accessed 20 June 2015
59. Wellner, B., Huyck, M., Mardis, S., Aberdeen, J., Morgan, A., Peshkin, L., Yeh, A., Hitzeman, J., Hirschman, L.: Rapidly retargetable approaches to de-identification in medical records. *J. Am. Med. Inform. Assoc.* **14**(5), 564–573 (2007)
60. Yeniterzi, R., Aberdeen, J., Bayer, S., Wellner, B., Hirschman, L., Malin, B.: Effects of personal identifier resynthesis on clinical text de-identification. *J. Am. Med. Inform. Assoc.* **17**(2), 159–168 (2010)

Chapter 27

Challenges in Synthesizing Surrogate PHI in Narrative EMRs

Amber Stubbs, Özlem Uzuner, Christopher Kotfila, Ira Goldstein,
and Peter Szolovits

Abstract Preparing narrative medical records for use outside of their originating institutions requires that protected health information (PHI) be removed from the records. If researchers intend to use these records for natural language processing, then preparing the medical documents requires two steps: (1) identifying the PHI and (2) replacing the PHI with realistic surrogates. In this chapter we discuss the challenges associated with generating these realistic surrogates and describe the algorithms we used to prepare the 2014 i2b2/UTHealth shared task corpus for distribution and use in a natural language processing task focused on de-identification.

27.1 Introduction

Before researchers can use electronic medical records (EMRs) outside of the institution that generated the records, it is critical they remove all protected health information (PHI) from the documents. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) [12] provides “safe harbor” guidelines which define what information is considered PHI. The list of PHI includes information such as patient names, ID numbers, phone numbers, email addresses, and

A. Stubbs (✉)

School of Library and Information Science, Simmons College, Boston, MA, USA
e-mail: stubbs@simmons.edu

Ö. Uzuner • C. Kotfila

State University of New York, Albany, NY, USA
e-mail: ouzuner@albany.edu; ckotfila@albany.edu

I. Goldstein

Department of Computer Science, Siena College, Loudonville, NY, USA
e-mail: igoldstein@siena.edu

P. Szolovits

Department of Computer Science and Engineering, Massachusetts Institute
of Technology, Cambridge, MA, USA
e-mail: psz@mit.edu

Admission Date : 06/07/99	Admission Date : 06/10/72
Discharge Date : 06/13/1999	Discharge Date : 06/16/2072
HISTORY OF PRESENT ILLNESS :	HISTORY OF PRESENT ILLNESS :
Mr. John Smith is a 60 year old male was last admitted in early January to Massachusetts General with chest pain. He was attended by Dr. Burke , then later treated by Dr. Amy Pagan and Luke Strauss, RN.	Mr. Thomas Benson is a 60 year old male was last admitted in early January to Orlando Regional with chest pain. He was attended by Dr. Newcomb , then later treated by Dr. Yuridia Joy and Milo Brock , RN.
Smith is a bartender who primarily works at Publick House , but sometimes works weekdays at Cheers .	Benson is a Food Service Manager who primarily works at Ideal Industries , but sometimes works weekdays at JITB .

Fig. 27.1 A fabricated sample medical record before and after surrogate generation

so on. We refer to the process of removing PHI from EMRs as “de-identification”. Researchers tasked with de-identifying EMRs face two challenges: (1) Identifying all the PHI in a medical record, and (2) replacing the PHI with some type of placeholder or surrogate.

In charts and other structured medical data, it may be sufficient to replace a patient’s name with a generic placeholder such as “[**PATIENT NAME 12345**]”. However, in data that is intended to be read naturally, such as hospital discharge summaries and correspondence between doctors, it is important that the de-identification process maintains the discourse structure and readability of the original text. This readability is important both for humans who may want to make use of the text, but also for natural language processing systems that may use the text for training.

We refer to replacing identified PHI with natural-sounding replacement data as “surrogate generation”, though it is sometimes referred to as “re-identification”¹ [23]. We maintain that the surrogate generation process should be implemented in such a way that the replacement data retain the same forms as the original and, as much as possible, the same internal temporal and co-reference relationships.

Figure 27.1 shows a sample record that has gone through the surrogate generation process. The left side shows the “original” text (a medical record fabricated for this chapter); the right side shows the processed record.

In this chapter we focus on the task of surrogate generation by discussing related work (Sect. 27.2), the HIPAA definitions of PHI and whether they are sufficient for

¹Somewhat confusingly, “re-identification” is also sometimes used to refer to determining a person’s true identity from de-identified data [7], so we avoid that term for the remainder of this chapter.

true de-identification (Sect. 27.3), the data used for this case study (Sect. 27.4), and the difficulties associated with generating realistic surrogate PHI in clinical data and the approaches we took in generating surrogate PHI for the 2014 i2b2²/UTHealth³ natural language processing (NLP) shared tasks (Sect. 27.5) [25, 30]. Finally, we discuss potential errors introduced by the surrogate generation process (Sect. 27.6) and the relationship between de-identification and surrogate generation (Sect. 27.7).

27.2 Related Work

Studies have shown that the majority of the U.S. population can be identified using only their gender, date of birth, and ZIP code [10, 28]. Another recent study found similar results for Canadian residents using records of people’s addresses over an 11 year period [7]. If a patient’s true identity were to be reconstructed from poorly de-identified medical records, that breach of trust could potentially endanger future research that relies on sharing de-identified medical records.

In order to test the efficacy of the 18 HIPAA categories for protecting patient identities, Lafky et al. [17] performed a study to determine how easily de-identified records could be linked back to the patient’s true identity. They obtained (with IRB approval) a set of approximately 15,000 de-identified patient records. These patients were all from a minority ethnic group, which the researchers thought would make the identification process easier. They then obtained a list of individuals from a commercial data repository; the list was matched to the ethnic group and geographic area (the specifics of the information in this list were not discussed). They identified unique patient records, and then matched them against the commercial data. Once they thought a match was found, another team verified their guess against the original records. In the end, only two patients of the 15,000 were correctly identified. Lafky’s research demonstrates that eliminating or pseudonymizing the 18 HIPAA categories of PHI makes the chance of determining patients’ true identities very low, although not entirely impossible.

In a more recent study, Meystre et al. [22] explored whether doctors could recognize their own patient’s de-identified notes. The authors used their own automated system to remove PHI from recent clinical notes (1–3 months old), then asked physicians to try to identify their patients. None of the doctors correctly identified their patients.

These studies show that good recognition of PHI is critical for patient protection, and indeed, much of the research in the area of EMR de-identification is in building systems that locate and categorize PHI in EMRs. As Li et al. [19] recently noted, Conditional Random Field algorithms (CRFs) [16] are a prominent approach to finding PHI and similar tasks, such as named entity recognition. Gardner and Xiong

²Informatics for Integrating Biology and the Bedside.

³University of Texas Health Science Center at Houston.

[8] report similar findings, and this observation was reinforced during the 2014 i2b2/UTHealth NLP shared task. This task featured a track on de-identifying clinical narratives [25, 30], and of the top ten systems, five of them used CRFs to identify PHI [24].

However, a CRF trained on gold-standard PHI data is not necessarily enough for complete de-identification, as the results from the 2014 i2b2/UTHealth shared task show [24], and researchers have been looking at other methods to improve automated systems. Interesting recent approaches include augmenting models with data from public medical texts [20] and clustering clinical narratives by complexity prior to de-identification [19].

While PHI identification is a critical component of the surrogate generation process, the focus of this chapter remains on surrogate generation itself. We refer readers interested in de-identification systems to two recent review articles on the subject: Kushida et al. [15] and Meystre et al. [21], as well as Chap. 26 in this volume.

For the remainder of this chapter we focus our attention on surrogate generation systems and procedures that researchers have used to de-identify authentic narrative EMRs to make the records available for researchers outside their home institutions. As we noted earlier, some datasets make use of generic placeholders or other forms of obfuscation when removing PHI [1, 2, 8, 11, 13]. This chapter focuses on surrogate generation techniques that preserve the readability of the natural language found in the original documents, so the remainder of this section predominantly examines surrogate generation systems in that paradigm, and the methods they used.

One of the first surrogate generation systems was Scrub [27], a system that matched the format of the replacement text to the original text. For dates, Scrub would group the detected dates to a point such as the first of the closest month. For names, Scrub used a look-up table, so that the same name in the original file would always be replaced with the same surrogate.

In creating the MIMIC II Clinical database, which is part of PhysioNet [9], Clifford et al. [3] used a two-step approach. First, they used the system created by Neamatullah et al. [23] to identify the PHI. Next, Neamatullah et al. describe generating realistic surrogate PHI to test their de-identification system, but the final version of the full MIMIC II corpus “scrubbed” the PHI by replacing the identified PHI with placeholders in the format “[** (data) **]”. “(data)” represents either a date-shifted piece of temporal information, or a marker to indicate what type of PHI it replaced, such as “First name 213” or “Hospital 57”. Each set of patient records (both narrative and tabular) has its own randomly-assigned date shift.

Douglass et al. [6] annotated PHI by hand and replaced them with realistic surrogates in a set of 2646 nursing notes from MIMIC II. The authors state that they shifted dates by a random number of weeks and years, while preserving days of the week. They generated names by mixing and matching from a list of Boston residents, and replaced locations with randomly selected small towns or, in the case of hospitals and wards, with information about a fictitious hospital. They maintained co-reference by making sure repeated mentions of the same authentic PHI were

replaced by the same surrogate PHI. At the end of surrogate generation, a human reviewed the suggested surrogates and had the option to modify the surrogate PHI to ensure that it was reasonable.

Uzuner et al. [29] used an earlier version of the realistic surrogate generation methodology we describe in this chapter, and one similar to that used by Douglass et al. [5]. Specifically, for strings such as ID numbers and phone numbers, they replaced the existing digits and letters with randomly-generated ones. For dates they retained the relations between times by offsetting all dates in a record by the same number of days, and ensured that the surrogate dates were properly formatted. For names of people and places, they randomly generated names by selecting syllables from existing names and mixing them together. Finally, to deliberately introduce ambiguity into their surrogates, they replaced some of the randomly-generated person names with medical terms, such as disease names and interventions. We modified the system used by Uzuner et al. [29] to create the system described in this chapter.

Deleger et al. [4] also used realistic surrogates. They generated surrogate names by randomly selecting male, female, unisex, and surnames from lists compiled from the US Census Bureau. They obfuscated phone numbers, ID numbers, and email addresses by randomly selecting new digits or letters as needed. Deleger et al. took a somewhat different approach to generating dates and locations than the other surrogate generation projects we have discussed so far. Rather than date-shifting by some amount of time or randomly selecting locations from a pre-existing list, they used the PHI in the corpus itself to compile lists of the different types of locations (streets, cities, etc) and different date formats (“November 2, 2013”, “4/27/03”). Then, they shuffled dates and parts of locations between documents in the corpus.

Removing all PHI from a record and replacing it with surrogate information protects patient privacy, but surrogate PHI is not always perfect. Yeniterzi et al., [31] recently compared the performances of a machine learning-based de-identification tool when trained and tested on different combinations of surrogate and authentic (the authors refer to it as “original”) PHI. They found that when they trained the system on surrogate (“resynthesized”) PHI, the system did not perform as well on authentic PHI because of the regularization imparted by the surrogate generation process.

In fact, the resynthesized-to-authentic model had the lowest performance of the four comparisons they performed, with f-measures ranging from 0.47 to 0.81. Authentic-to-authentic and resynthesized-to-resynthesized both did well (0.93–1.00 and 0.96–0.99, respectively), with the third-best performance on authentic-to-resynthesized (0.78–0.89). The fact that the systems performed best when trained on similar data suggests that, in order to provide the best out-of-the-box de-identification on authentic records, de-id systems must either be trained on authentic data (which is rarely possible) or that the resynthesized data must mimic natural language as closely as possible.

Using realistic surrogates maintains the discourse structure of the documents and helps train machine learning systems for de-identification of authentic PHI. However, most papers that discuss systems that generate surrogate PHI dedicate

only a paragraph or two to the surrogate generation process. In this chapter, we present a solution that generates realistic surrogate PHI while maintaining both coreference chains and temporal continuity within the narrative, and benefits from human oversight for best results. The remainder of this chapter discusses our surrogate generation method in detail, including the algorithms we used for certain types of PHI, the difficulties posed by the narrative texts, and the need for human intervention in the surrogate generation process. We hope that by creating better surrogate PHI for publicly-available data sets that we can improve the performance of de-identification systems designed on these data even on authentic medical records.

27.3 PHI Categories

HIPAA defines 18 categories of PHI of “the [patients] or of relatives, employers, or household members of the [patients],” that must be removed from medical records.⁴ These categories are described in Table 27.1.

As written, categories 1–17 do not include the names of doctors or other medical personnel, names or locations of hospitals or medical facilities, or any other information that identifies a person who is not a patient or directly related to the patient.

However, in our experience with de-identifying medical records, and based on the aforementioned studies on determining the real identity of individuals with minimal information [7, 10, 28], we have found it beneficial to adopt a risk-averse policy towards de-identification and surrogate generation. Our goal is to ensure patients’ privacy and to protect their identities, therefore it behooves us to make our best effort to remove any of the information in a record that a malicious person could use to identify a patient.

Consider the situation where an EMR mentions that a patient sees “Dr. Lau of Belfast Hospital”, but upon investigation one can learn that not only does Dr. Lau only see a few patients a year, she only treats patients with paranoid schizophrenia, and she was only at Belfast Hospital from 2001–2002. In this case, information about the doctor and the hospital, when triangulated with external knowledge from other sources, could lead to patient identification.

To minimize such risks, we expand our definition of the HIPAA category 18 to include the following:

- All person names in a document, including hospital staff and their user names;
- All locations, including states, countries, geographical areas (e.g., “The Northeast”), landmarks (e.g., “the Grand Canyon”), and non-generic hospital departments;
- Organizations (e.g., Simmons College, Google);
- All portions of dates, including years;

⁴45 CFR 164.514.

Table 27.1 HIPAA's list of PHI categories

-
1. Names;
 2. All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:
 - a. The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and
 - b. The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
 3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
 4. Telephone numbers;
 5. Fax numbers;
 6. Electronic mail addresses;
 7. Social security numbers;
 8. Medical record numbers;
 9. Health plan beneficiary numbers;
 10. Account numbers;
 11. Certificate/license numbers;
 12. Vehicle identifiers and serial numbers, including license plate numbers;
 13. Device identifiers and serial numbers;
 14. Web Universal Resource Locators (URLs);
 15. Internet Protocol (IP) address numbers;
 16. Biometric identifiers, including finger and voice prints;
 17. Full face photographic images and any comparable images;
 18. Any other unique identifying number, characteristic, or code.
-

- Pager numbers;
- Any of the PHI types included in categories 4–17 that apply to people or organizations other than the patient or the patient's associates;
- Professions held by the patient or people associated with the patient (this does not include the professions of the hospital staff);
- Any other information that potentially indicates a patient's location or a specific time in the document, such as references to historic events.

Figure 27.2 shows a fabricated medical record before surrogate generation; Fig. 27.3 shows the same record after surrogate generation.

We generate surrogates for PHI in all of the HIPAA categories, including category 18. The generated surrogates maintain co-reference between named entities (e.g., people, locations) by replacing authentic PHI that occurs more than once with the same surrogate, and they maintain temporal relationships by shifting all the dates in a patient's records forward by the same interval. We discuss specific details of the surrogate generation process in the following sections, as well as the issues surrounding generating appropriate surrogate information for all the PHI categories, and the effect these surrogates have on certain types of medical research.

```

Record date: <DATE>2011-01-14</DATE>

INFECTIOUS DISEASE ASSOCIATES
<HOSPITAL>CURTIS MEDICAL CENTER</HOSPITAL>

Patient: <PATIENT>Iles, Lois</PATIENT>
Attending: <DOCTOR>Riley, Todd M.</DOCTOR>

<AGE>70</AGE>y/o F was seen in ID clinic on
<DATE>7/20</DATE>, <DATE>9/27</DATE>, <DATE>11/9</DATE>,
and today (<DATE>1/14/11</DATE>).

PMH: DMII (dx late <DATE>90s</DATE>, last A1c = 5.6 in
<DATE>2/10</DATE>)

[...]
Ms. <PATIENT>Iles</PATIENT> was seen and examined with Dr.
<DOCTOR>Tillman</DOCTOR>.

<DOCTOR>Todd Riley</DOCTOR>, MD pager #<PHONE>03268</PHONE>
cc: <DOCTOR>DANIELLE TOMPKINS</DOCTOR>,
<HOSPITAL>CMC</HOSPITAL> Internal Medicine
Signed electronically by
<DOCTOR>Todd Riley</DOCTOR>, MD
<USERNAME>TMR42</USERNAME>

```

Fig. 27.2 Fabricated EMR prior to surrogate generation; PHI are delineated with XML tags

27.4 Data

We developed and tested the surrogate generation algorithms described here over years on two data sets: the 2006 and 2014 i2b2/UTHealth shared task corpora. The 2006 corpus consists of 889 medical discharge summaries [29], which are narrative descriptions of a patient's hospital stay, written at the time of discharge. The corpus contains 19,498 instances of PHI from eight PHI categories: patients, doctors, locations, hospitals, dates, IDs, phone numbers, and ages [29].

The 2014 i2b2/UTHealth shared task corpus is made up of 1304 longitudinal medical records (records that refer to the same patient over a period of time) for 296 patients [14]. In addition to discharge summaries, the 2014 corpus contains other types of narrative medical records, such as inpatient notes and correspondence between specialists and primary care physicians. The 2014 i2b2/UTHealth corpus contains 28,872 instances of PHI [25], and uses the expanded list of PHI categories we described in Sect. 27.3.

The 2006 corpus provided the original data on which the surrogate generation algorithms were tested; we then refined the algorithms for use with the 2014 corpus, as it contained a wider variety of both record types and PHI. The rest of this chapter focuses on how we applied surrogate generation to the 2014 data set; a discussion of the surrogate generation process on the 2006 data can be found in Uzuner et al. [29].

Both the 2006 and 2014 corpus files come from Partners Healthcare, and the Institutional Review Boards (IRBs) of Partners Healthcare, Massachusetts Institute of Technology, and the State University of New York at Albany approved the data

```
Record date: <DATE>2094-02-06</DATE>

INFECTIOUS DISEASE ASSOCIATES
<HOSPITAL>HARRISON COUNTY INFIRMARY</HOSPITAL>

Patient: <PATIENT>Kuhn, Naomi</PATIENT>
Attending: <DOCTOR>Robertson, Carlos U.</DOCTOR>

<AGE>70</AGE>y/o F was seen in ID clinic on
<DATE>8/12</DATE>, <DATE>10/20</DATE>, <DATE>12/2</DATE>,
and today (<DATE>2/06/94</DATE>).
PMH: DMII (dx late <DATE>70s</DATE>, last A1c = 5.6 in
<DATE>3/93</PHI>)
[...]
Ms. <PATIENT>Kuhn</PATIENT> was seen and examined with Dr.
<DOCTOR>Chung</DOCTOR>.
<DOCTOR>Carlos Robertson</DOCTOR>, MD pager #<PHONE>86962</PHONE>
cc: <DOCTOR>FAYE COX</DOCTOR>,
<HOSPITAL>HCI</HOSPITAL> Internal Medicine
Signed electronically by
<DOCTOR>Carlos Robertson</DOCTOR>, MD
<USERNAME>CUR25</USERNAME>
```

Fig. 27.3 Fabricated EMR after surrogate generation; PHI are delineated with XML tags

preparation we describe here, as well as the release of this data under a Data Use Agreement (DUA). The 2014 corpus data will be made available to researchers in November 2015 from <http://i2b2.org/NLP>; the 2006 data is already available at the same location.

27.5 Strategies and Difficulties in Surrogate PHI Generation

For many of the PHI categories, generating realistic surrogates is relatively simple. For example, phone and fax numbers, URLs, email addresses, and ID or account numbers (i.e., HIPAA categories 4–16), whether they are for patients or medical staff, are usually alphanumeric strings of numbers, letters, and a few special characters such as hyphens, parentheses, periods, and the ‘at’ character (@). In these cases, it is simple to not only replace the authentic PHI with randomly-generated surrogates, but it is also relatively easy to maintain co-references.

Figure 27.4 presents a summary of the algorithm we used to replace PHI represented by numeric or alphanumeric sequences. By keeping track of the category of each PHI, we were able to maintain co-reference between elements of the same category without revealing similarities between PHI of different categories. For example, given a document with the phone numbers (555) 123–4567 and 123–4567, as well as the medical record ID number 1234567, we can easily maintain co-reference relationships between the two phone numbers, but we would

1. Isolate the string of numbers and digits; check to see if that string or a substring already has a replacement in the appropriate PHI category.
2. Replace any digit by a randomly-generated digit.
3. Replace any lower-case or upper-case alphabetic character by a randomly-generated lower-case or upper-case, respectively, alphabetic character.
4. Leave other characters and spacing alone.

Fig. 27.4 Generic algorithm for replacing alphanumeric strings

randomly replace the medical record ID digits without referring to the already-replaced phone PHI.

Naturally, we implement more specific rules for various categories of PHI. For example, phone numbers have a restriction that they cannot start with 0, and the substring rule does not apply to medical record numbers or other account or ID numbers. This replacement algorithm resulted in unrealistic email addresses and URLs, however. For example: “rgs44@newhospital.org” might be replaced with “bhg10@jrbsotnwwx.hrh”. Given the small number of email addresses and URLs in the corpus, we generated surrogates for these by hand, though later implementations of this system may involve more complex solutions to this problem, involving matching pieces of addresses with names and hospitals found in the text.

While the algorithm in Fig. 27.4 is relatively simple, not all co-referent PHI are so easily replaced. The remainder of this section addresses some of the challenges we faced with replacing more complex PHI, as well as the approaches we took to address them. We present these PHI in order of HIPAA categories that most closely match them. Because we are discussing only text-based files, HIPAA categories 16 and 17 (biometrics and facial photographs) are not included in this discussion.

One challenging aspect of the 2014 i2b2/UTHealth NLP shared task corpus that augmented the complexity of generating surrogate PHI is the longitudinal nature of the records in the corpus. We mitigated the problem of cross-document co-reference by merging all of the records for a single patient into a single file for the purposes of surrogate generation. We then processed all of these records together, thereby ensuring that co-references would not be lost between the patient’s records. However, having multiple records per patient increased the potential for variations and misspellings of co-referent PHI that our surrogate generator had to handle.

27.5.1 HIPAA Category 1: Names

As previously noted, the PHI in our corpus include the names of medical professionals in addition to patients and their relations. Names provide an interesting challenge in surrogate generation because they can, and do, occur in many different forms within the same document. Also, as previously noted, increasing the number

of documents about a single patient correspondingly increases the number of co-referent PHI. Over the course of multiple documents, a single patient could, for example, be referred to as:

- Vasquez, Angela
- Angela Vasquez
- Angie
- Ms. Vasquez
- Angela M. Vasquez
- and various misspellings of these names, such as “Angel” or “Vazquez”.

Nicknames and misspellings are particularly tricky for an automated system to catch. While it is likely that “Angie Vazquez” and “Angela Vasquez” are the same person, it is also possible for these two names to refer to two different people, e.g., a patient and a doctor, something that human readers may be able to distinguish but a simple surrogate generation system may not.

Another layer of challenge is that, in the i2b2 corpora, doctors often initial the bottom of records to indicate they are complete, and sometimes add their hospital system username after their own name. This means that in addition to maintaining co-reference between full names, we needed to be able to identify initials and map them appropriately to the full names, while differentiating them from acronyms that may look exactly the same as these initials.

Finally, we wanted to maintain the genders of the names in the original documents as much as possible. To help with this task, we obtained information from the US Census Bureau and split the information into four dictionaries for lookup: (1) surnames (2) female names (3) male names (4) unisex names. With this resource, we approached surrogate name generation with the algorithm summarized in Fig. 27.5.

Pre-mapping the letters of the alphabet to letters that we then used to select surrogates had many advantages. First, it allowed us to deal easily with ambiguous co-references in the text without having to make leaps of inference. For example, Fig. 27.2 refers to Drs. Tillman and Thompkins. If at some point the narrative also referred simply to “Dr. T.”, our system would not have to attempt to disambiguate the reference. Since “Tilman” and “Tompkins” are both replaced with names starting with the same letter (“C” in Fig. 27.3), the ambiguity in the original document is maintained in the surrogate as well. Similarly, the alphabetic mappings allowed us to easily replace initials and usernames without having to infer which doctor might have signed the document.

While overall this algorithm worked fairly well, there were a few problems with certain circumstances. First, we lacked sufficient support for nicknames and misspellings, and so this type of co-reference had to be corrected by hand. Future implementations will utilize a nickname dictionary as well as some rules involving Levenshtein distances [18], which will help identify small errors or misspellings. Second, randomly creating mappings for letters sometimes led to situations where a letter that occurs commonly in names, such as S, would be mapped to a letter that is much less common, such as X, meant that our dictionary would sometimes be

1. Given a record or set of records for a single patient, create two random mappings between letters in the alphabet: one for given names (first, middle) and one for last names (i.e., A mapped to R, B mapped to E, etc.).
2. For each PHI in the name category, normalize the format into categories that correspond with first name, middle name(s), surname, and suffix.
 - a. Check to see if any of these PHI have already been mapped to surrogates.
 - b. If they have, use those mappings.
 - c. If they have not:
 - i. Check against the name dictionaries to determine if the name is most likely male, female, unisex, or a surname.
 - ii. Use the alphabet mapping for the appropriate name type to randomly select a new surrogate. For example, if “Angela” is determined to be female and A is mapped to R for this document, select a new name from the female dictionary that begins with R.
3. If the name cannot be normalized, assume it is a set of initials and replace them based on the alphabetic mappings. If the name is a username (initials followed by numbers), map the initials and generate random surrogate numbers.
4. Place the new PHI/surrogate mappings in a lookup table.

Fig. 27.5 Algorithm for generating replacement names

forced to re-use the same surrogates for different PHI. This lack of names starting with certain letters potentially added co-references and/or ambiguity where none existed in the original file. One solution to this would be to set the letter to letter mappings to roughly follow distributions of the census data, though this approach could potentially lead to a PHI leak, as it would reveal information about the authentic PHI.

Finally, randomly selecting first, middle, and last names from dictionaries sometimes leads to extremely improbable surrogate names, especially if the system fails to accurately categorize a name as a given name or a surname. One notable example from an early test of our surrogate system led to the creation of a doctor named “Pagan Trout”. Unfortunately, Dr. Trout was not included in the final i2b2/UTHealth corpus.

27.5.2 HIPAA Category 2: Locations

The different location types we addressed were: countries, states, ZIP codes, cities, streets, hospitals, organizations, hospital departments, room numbers, and other locations, such as landmarks and geographic regions.

With the exception of ZIP codes, which we changed by using the algorithm in Fig. 27.4 to generate surrogate PHI for locations we created pre-compiled lists of different location types, and randomly selected from those when generating surrogates. However, we had to implement special rules for certain location types. For example, it is common for hospital names to take many different forms in even

a single medical record. “Massachusetts General Hospital” might be referred to by its full name, as “Mass. General”, as “Mass. Gen.” or simply as “MGH”. Therefore, the algorithm for hospital name replacement looked for possible abbreviations and modified the surrogate output to match the format of the original PHI.

We included hospital department names in our list of PHI in order to ensure that a hospital could not be identified by a department name that is unique. We made department names less identifiable by mapping them to a list of generic department names. For example, the department of “anesthesia, critical care, and pain medicine” would become “anesthesiology”. We left alone department names that were already generic (i.e., “Emergency Department”, “Oncology”). Any department that could not be mapped to a more generic version we adjusted by hand.

We also needed to correct demonym (name for persons from a country or other location) by hand, as the software did not check for them. If a person were described as “from Armenia” in one EMR and “Armenian” in another, we would check to make sure that the surrogate country of origin was the same in both EMRs, and that the demonym took the appropriate form.

27.5.3 HIPAA Category 3: Dates and Ages

One of the tasks in the 2014 i2b2/UTHealth challenge involved a temporal analysis of each patient’s medical records, and so we prioritized maintaining the temporal relationships between all the dates in a patient’s set of records while still obfuscating the authentic dates in the record. We did this by shifting all the dates in a record into the future by the same interval, but randomly selecting the interval for each new patient. Thus we were able to maintain continuity for each patient without revealing any identifiable dates.

In order to shift all the dates in a patient’s records consistently, we first had to identify all of the dates in the records and convert them to a standard format. While co-reference is not a problem with dates in the way that it is with names, converting dates to a standard format still poses a challenge. A date such as “September 29, 2013” is easy to parse and standardize to 2013-09-29, but many other formats are more difficult to interpret. For example, if the same date were represented as 09/29/2013 we could easily tell that the format is mm/dd/yyyy, as “2013” is not a valid month or day, and “29” is not a valid month. However, the date 11/10/13 is ambiguous. While the “13” represents the year in most date formats, whether the “11” represents the 11 month or the 11th day is unclear. For our system, we first introduced logical constraints: a month cannot be a number greater than 12, and a day cannot be greater than 31. In cases where this logic did not help, such as “11/10/13”, we assumed a mm/dd/yy representation, which is more common in America where these records originated. In order to interpret the year, we assumed that any two-digit year of 20 or less applied to the twentyfirst century, while years ending in 21 through 99 applied to the twentieth century.

An ambiguity that we found less easy to resolve was that of two-integer dates, such as “04/03”. Even if we assume the American convention that this date represents April 3rd rather than March 4th, we are still left with the possibility that the date could represent April of 2003. Our manual review of these dates in the original texts revealed that both uses of that format (i.e., mm/dd or mm/yy) occur in the patient records, and we could not implement a rule to automatically shift these dates. In the end, our system marked these dates as ambiguous and a human had to interpret and shift them manually.

Other dates, such as named holidays (i.e., Christmas, Halloween) we mapped to calendar dates whenever possible. Any date that could not be converted into a standard format we marked as “unknown” and left to human interpretation. Most often, these were cases where the dates were malformed, such as “4/301999”, which is missing a/between the day and the year.

Once the system converted all of the dates that it could to a standard format, the system performed a check on the span of identified dates, to make sure that the difference between the earliest and latest date is less than 90 years. This check is motivated by the HIPAA requirement that ages over 89 be obfuscated. If the ages in a document are already annotated, identifying the ones that are over 89 is trivial; our system changed all the ages over 90 to “90” and left the other ages alone. However, if a patient is over 89 and the file lists their birth date and the current date, it would be easy to infer their actual age if all the dates are shifted consistently. Therefore, if the span of dates in a document is over 90 years, the system identified the earliest dates and shifted them more years into the future than the other dates in the records.

Finally, the system randomly selected the number of years and days that it would use to shift the dates in the record. We limited the number of years that dates could be shifted forward to 65 years plus or minus a maximum of 20 years for the 2014 i2b2/UTHealth corpus. The number of days we shifted the dates could be positive or negative, and we calculated this number by determining the maximum number of days the dates could be shifted in a direction without changing the season of any of the dates, then selecting a random number between one and the maximum. The day shift also acted as the month shift: if we shifted all the days back by seven, October 1st would become September 24th.

However, this method of date shifting also introduces its own potential sources for error. Consider a medical record that contains the statement “he will schedule a follow-up for January”. The medical implication of this phrase is changed drastically if the record is dated December 24th (implying that the patient needs to be seen mere weeks, or even days, after that visit) or January 3rd (implying that the patient is anticipated to be fine for another year). If our date-shifting algorithm keeps the date on the record to be sometime in December, then “January” can be unchanged without drastically changing the implied medical condition of the patient. However, if “December 24th” becomes “January 9th”, and the mention of the “follow-up in January” remains unchanged, then the patient’s implied condition goes from being in need of close monitoring to safe for another year.

Any isolated mention of a month has the same problem. Our system addressed this problem by assigning a hidden day-of-the-month variable to the unanchored

months and giving that variable the value of “15”. Then it applied the normal date-shifting algorithm, and the month would shift to the next one if the day-shift pushed it into the next month. So, given the above example with “December 24” and the following “January”, if the date-shift value was set to 17 then the new dates would be “January 8” and “February”, because the “hidden” 15th of January would become February 1st. If the date-shift was less than 17, then January would remain January. However, any forward date-shift less than seven would mean that the December date would remain in December, so keeping “January” unchanged would be the appropriate action.

Our method for dealing with unanchored months is not foolproof and can add errors into the surrogate text. For example, if the original dates were “August 2” and “the following September”, if the dates were shifted forward by 16 days, the results would be “August 18” and “the following October”, due to the hidden variable “15”. An ideal approach would be to implement a system that uses the surrounding context to infer the year in which the unanchored month took place. Unfortunately, the task of assigning calendar dates to temporal phrases in medical records is no mean feat. The 2012 i2b2 NLP shared task focused on temporality in clinical texts, and the best-performing system had a 0.9 f-measure for identifying temporal expressions, but only achieved 0.73 for accuracy of the attribute values, which included the day, month, and year for identified dates [26].

27.5.4 HIPAA Category 18: Other Potential Identifiers

Additionally, we had to address information in the documents that did not fall into the PHI categories we described above, but that still had the potential to compromise information about a patient. Some EMRs contain references to specific events that the patient took part in, such as “injured during Superstorm Sandy” or “enrolled in HEART-FAB study”. This type of information we usually removed entirely, or modified to be less identifiable. “Injured during Superstorm Sandy” could have become “Injured during last week’s thunderstorm”.

27.5.4.1 Professions

As previously discussed, we included patient’s professions in the expanded set of PHI categories. The algorithm we used to generate surrogates for professions was relatively simple: we either selected a new, random profession from a pre-compiled list, or we used a surrogate that had already been selected for the same PHI. From a software perspective, this was easy to implement, but viewing professions as PHI posed a challenge that ultimately had to be handled through human intervention.

A person’s profession can influence their health risks and outcomes. For example, people whose occupations involve construction or other building-related work, such as electricians and plumbers, are more likely to be exposed to asbestos,

and therefore those workers are more at risk for mesothelioma than others.⁵ The relationship between medical outcomes and professions led us to attempt to assign new professions with roughly the same types of medical risks, thus preserving anonymity without entirely losing potential medical causalities. Similarly, the job held by a patient's relative occasionally impacted the medical record. For example, a record might state "Patient's daughter is a registered nurse who will help him monitor his blood pressure and blood sugar." Changing "registered nurse" to "lawyer" or another non-medical profession would make the sentence less logical, while supplying "doctor" or "medical aide" as a replacement would maintain the logic of the narrative. Our surrogate system did not implement a job hierarchy, and so we did this portion of the surrogate generation by hand where necessary.

27.6 Errors Introduced by Surrogate PHI

We made every effort to maintain the continuity of patient's medical histories so that researchers could still use the files for medical NLP. However our prioritization of patient privacy required us to make decisions that fundamentally changed the nature of the corpus.

By shifting each patient's records by different intervals and by randomizing the locations in the text, we rendered the documents useless for epidemiological studies, as researchers cannot use the data to infer trends based on shared locations and points in time. Errors in date-shifting can lead to creating mistakes in patients' own timelines, and if the system mistakenly generates different surrogates for entities that are co-referent, information about the patient is again lost. Similarly, substituting professions, even ones of similar types, may remove relevant job-related risk factors. Finally, any modifications to a text can result in unrealistic narratives in the form of incorrect determiners, verb tenses, and other grammatical errors.

Having a human review and edit the output can help reduce some of these errors, such as lost co-references and grammatical issues, but this is a time-consuming process and may not be possible for most research teams focused on de-identification.

27.7 Relationship Between De-identification and Surrogate Generation

The task of a de-identification algorithm should be to identify PHI, not to obliterate it. This is because obliteration (e.g., by replacement by a marker such as "[**

⁵<http://www.asbestos.com/occupations/>.

NAME 33 **]”) destroys information about the original format of that PHI. This information is crucial for generating realistic surrogates.

Furthermore, a surrogate generation system may be able to do a better job in generating surrogates if it has access to the features, patterns, and dictionary memberships that the de-identification system used to detect the PHI. Otherwise, it may need to re-do some of the work already accomplished by the de-identifier. Our experience also shows that determining accurate co-references in the original data would be a good step toward creating consistent surrogates. We have described some fairly ad hoc methods for doing this, but a more general purpose method that itself uses machine learning techniques to build models that identify co-reference might be more effective. It does seem that, contrary to our initial expectations, it is not necessarily optimal to factor de-identification and surrogate generation into separate sequential tasks, connected only by a narrow stream of PHI annotations.

27.8 Conclusion

In this chapter we have presented some methods and algorithms for approaching the task of realistic surrogate generation for narrative EMRs, and discuss some of the specific approaches we used when creating the de-identified dataset for the 2014 i2b2/UTHealth NLP shared tasks.

In the United States, HIPAA regulations require that researchers can only release medical records for research purposes if each patient has given consent or if the records have been de-identified in order to protect patient privacy. By removing authentic PHI and replacing them with realistic surrogates, we attempt to maintain the narrative structure of the original records. Our method for surrogate generation tackles PHI categories one at a time, and follows different mechanisms for generating appropriate surrogates for each kind. It respects co-reference, maintains ambiguities that naturally exist in the original data, but does not resolve them. It aims to minimize the ambiguities that appear as artifacts of the process. Our experiences show that while automated methods provide a good start at surrogate generation, manual review and human intervention makes surrogates natural. Nonetheless, a general method for surrogate generation comes with a cost—it can invalidate the data for some purposes. In our case, the date shifting choices make epidemiology research on this data untenable.

Future work in this area will need to focus on both identifying text that falls under HIPAA category 18 and developing reasonable surrogates for that information without removing important information about each patient’s health. A resource that groups occupations by similar health risks could potentially help with this problem. Improved systems for detecting co-reference, especially ones that can account for misspellings and nicknames, would also benefit this area of research.

Acknowledgements This project was funded by NIH NLM 2U54LM008748 PI: Isaac Kohane, and by NIH NLM 5R13LM011411 PI: Ozlem Uzuner.

References

1. Berman, J.J.: Concept-match medical data scrubbing. How pathology text can be used in research. *Arch. Pathol. Lab. Med.* **127**(6), 680–6 (2003)
2. Chakaravarthy, V.T., Gupta, H., Roy, P., Mohania, M.K.: Efficient techniques for document sanitization. In: Proceedings of the 17th ACM conference on Information and knowledge management, pp. 843–852 (2008)
3. Clifford, G.D., Scott, D.J., Villarroel, M.: User Guide and Documentation for the MIMIC II Database, database version 2.6. Available online: <https://mimic.physionet.org/UserGuide/UserGuide.html> (2012)
4. Deleger, L., Lingren, T., Ni, Y., Kaiser, M., Stoutenborough, L., Marsolo, K., Kouril, M., Molnar, K., Solti, I.: Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. *J. Biomed. Inform.* Aug;50:173–83 (2014). doi: 10.1016/j.jbi.2014.01.014
5. Douglass M.M.: Computer-assisted de-identification of free-text nursing notes. MEng thesis, Massachusetts Institute of Technology (2005)
6. Douglass M.M., Clifford, G.D., Reisner, A., Moody, G.B., Mark, R.G.: Computer-assisted deidentification of free text in the MIMIC II database. *Comput. Cardiol.* **31**, 341–344 (2004)
7. El Emam, K., Buckeridge, D., Tamblyn, R., Neisa, A., Jonker, E., Verma, A.: The re-identification risk of Canadians from longitudinal demographics. *BMC Med. Inform. Decis. Mak.* **11**, 46 (2011)
8. Gardner, J., Xiong, L.: An integrated framework for de-identifying unstructured medical data. *Data Knowl. Eng.* **68**(12), 1441–1451 (2009)
9. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E.: PhysioBank, PhysioToolkit, and Physionet: components of a new research resource for complex physiologic signals. *Circulation* **101**(23), e215–e220 (June 13, 2000). <http://circ.ahajournals.org/cgi/content/full/101/23/e215>
10. Golle, P.: Revisiting the uniqueness of simple demographics in the US population. In: Workshop on Privacy in the Electronic Society (2006)
11. Gupta, D., Saul, M., Gilbertson, J.: Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. *Am. J. Clin. Pathol.* **121**(2), 176–186 (2004)
12. HHS (Department of Health and Human Services). Standards for Privacy of Individually Identifiable Health Information, 45 CFR Parts 160 and 164. December 3, 2002 Revised April 3, 2003. Available from: <http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/introduction.html>
13. Jiang, W., Murugesan, M., Clifton, C., Si, L.: t-Plausibility: semantic preserving text sanitization. In: 2009 International Conference on Computational Science and Engineering (CSE), pp. 68–75 (2009). doi:10.1109/CSE.2009.353
14. Kumar, V., Stubbs, A., Shaw, S., Uzuner, O.: Creation of a new longitudinal corpus of clinical narratives. *J. Biomed. Inform.* 2015.
15. Kushida, C.A., Nichols, D.A., Jadrnicek, R., Miller, R., Walsh, J.K., Griffin, K.: Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med. Care* **50**, S82–S101 (2012)
16. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, pp. 282–289. Morgan Kaufmann, San Francisco (2001)
17. Lafky, D.: The Safe Harbor method of de-identification: an empirical test. Fourth National HIPAA Summit West. http://www.ehcca.com/presentations/HIPAAWest4/lafky_2.pdf (2010)
18. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR.* **163**(4), 845–848 (1965) [Russian]. English translation in Sov. Phys. Dokl. **10**(8), 707–710 (1966)

19. Li, M., Carrell, D., Aberdeen, J., Hirschman, L., Malin, B.: De-identification of clinical narratives through writing complexity measures. *Int. J. Med. Inform.* **83**(10), 750–767 (2014)
20. McMurry, A.J., Fitch, B., Savova, G., Kohane, I.S., Reis, B.Y.: Improved de-identification of physician notes through integrative modeling of both public and private medical text. *BMC Med. Inform. Decis. Mak.* **13**, 112 (2013). doi:10.1186/1472-6947-13-112
21. Meystre, S., Friedlin, F., South, B., Shen, S., Samore, M.: Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med. Res. Methodol.* **10**, 70 (2010)
22. Meystre, S., Shen, S., Hofmann, D., Gundlapalli, A.: Can physicians recognize their own patients in de-identified notes? *Stud. Health Technol. Inform. Stud Health Technol Inform.* 2014;205:778–82
23. Neamatullah, I., Douglass, M., Lehman, L.-W., Reisner, A., Villarroel, M., Long, W., Szolovits, P., Moody, G., Mark, R., Clifford, G.: Automated de-identification of free-text medical records. *BMC Med. Inform. Decis. Mak.* **8**, 32 (2008)
24. Stubbs, A., Kotfila, C., Uzuner, Ö.: Automated systems for the de-identification of longitudinal clinical narratives. *J Biomed Inform.* 2015 Jul 28. pii: S1532-0464(15)00117-3. doi: 10.1016/j.jbi.2015.06.007
25. Stubbs, A., Uzuner, Ö.: Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus *J Biomed Inform.* 2015 Aug 28. pii: S1532-0464(15)00182-3. doi: 10.1016/j.jbi.2015.07.020
26. Sun, W., Rumshishky, A., Uzuner, Ö.: Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J. Am. Med. Inform. Assoc.* Published Online First 5 April 2013
27. Sweeney, L.: Replacing personally-identifying information in medical records, the scrub system. In: Cimino, J.J. (ed.) *Proceedings, Journal of the American Medical Informatics Association*, pp. 333–337. Hanley and Belfus, Washington (1996)
28. Sweeney, L.: Uniqueness of Simple Demographics in the U.S. Population. Carnegie Mellon University, School of Computer Science, Data Privacy Laboratory, Technical Report LIDAP-WP4. Pittsburgh (2000)
29. Uzuner, Ö., Luo, Y., Szolovits, P.: Evaluating the state-of-the-art in automatic de-identification. *J. Am. Med. Inform. Assoc.* **14**(5), 550–563 (2007)
30. Uzuner, Ö., Stubbs, A., Xu, H., co-chairs.: “Data Release and Call for Participation: 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data”. <https://www.i2b2.org/NLP/HeartDisease/>
31. Yeniterzi, R., Aberdeen, J., Bayer, S., Wellner, B., Hirschman, L., Malin, B.: Effects of personal identifier resynthesis on clinical text de-identification. *J. Am. Med. Inform. Assoc.* **17**, 159–168 (2010)

Chapter 28

Building on Principles: The Case for Comprehensive, Proportionate Governance of Data Access

Kimberlyn M. McGrail, Kaitlyn Gutteridge, and Nancy L. Meagher

Abstract The amount of data in the world is growing rapidly. Researchers and others see the value of these data to answer compelling questions, and sometimes this involves linking different data sets together. Good and long-standing processes for governing access to data exist, but these will be challenged with the amount and breadth of data researchers wish to use. In particular, it is increasingly clear that in this new world of data, data access governance cannot continue to rely on traditional approaches of de-identification, anonymization and individual consent. An alternative to these risk-minimization approaches is proportionate governance, a process that assesses potential risks and mitigations to those risks, including the potential public interest that is served by enabling research. We propose a flexible and adaptable proportionate governance framework that builds on existing models. Local adoption of this framework will require engagement with stakeholder to create consensus around principles, and implies broad commitment to the notion of a more open research culture.

28.1 Introduction

One of the most notable features of the last 20 years is an increasing orientation to data [44, 45]. This orientation cuts across all sorts of areas of interest, applies equally in the public and private sector, and certainly applies to research from many academic disciplines. Researchers collect data on individuals and their habits and

K.M. McGrail (✉)

Centre for Health Services and Policy Research,
School of Population and Public Health, University of British Columbia,
Vancouver, BC, Canada

Population Data BC, University of British Columbia, Vancouver, BC, Canada
e-mail: Kim.mcgrail@ubc.ca

K. Gutteridge • N.L. Meagher

Population Data BC, University of British Columbia, Vancouver, BC, Canada
e-mail: kaitlyn.gutteridge@popdata.bc.ca; nancy.meagher@popdata.bc.ca

outcomes of interest. Clinicians collect data in order to track and care for their patients. Administrators collect data in order to pay for and manage the distribution and delivery of services. Cutting edge data storage technology 20 years ago was the 100 mB zip drive; today's average smartphone can store hundreds of times this, making both the collection and use of data easier for everyone involved.

The collection of data generally serves a primary purpose—such as your family doctor keeping track of your medical history to ensure you receive high quality care—but can have power well beyond this. Data can have significant utility in their secondary uses, defined simply as use for a purpose other than that articulated for their original collection [7]. Routinely collected health data can have a secondary purpose, for example, to assess the safety and effectiveness of drugs once they are approved for use in the general population [71].

Secondary use of existing health-related data is attractive largely due to efficiency, since the time-consuming and expensive process of data collection has already been completed. Beyond this, in the case of administrative data related to the running of or payment for health care services, the data are often population-based, which mitigates concerns about bias and power. Researchers have been accessing data for secondary purposes for decades, informing both health care practice and the development of health care and public policy [67].

The power of secondary data can be extended further if different sources are linked together. Data linkage makes it possible to identify (for example) the same individual as he or she moves from one part of the health care system to another [3, 9, 35]. Combining data from multiple sources is also a powerful way to expand the sorts of research questions that can be addressed and to improve robustness of results. This is because combining sources provides richer information on the individuals involved, which helps control for more potential confounders [64].

Assembly and governance are two distinct processes in the development of comprehensive linked data systems. Assembly is essentially the technical process of building capacity for data storage, linkage across different originating data sets, and access that may include processing. Data governance in contrast is more procedural than technical. Governance broadly speaking refers to the necessary components for achieving a collective goal among multiple stakeholders, and typically includes three elements. First, some sub-group or body is given authority to act on behalf of the group; second, a process or structure is set up to frame the decision-making boundaries for that group; and third, it is understood that there then must be formal accountability for those decisions [33]. Applied more specifically to data governance and even more directly to governance of access to data, governance “... refers to who holds the decision rights and is held accountable for an organization’s decision-making about its data assets” [37]. Data governance is more complex when it involves data from multiples sources and takes on particular importance in the realm of health because of the sensitive nature of the information involved.

Small pockets of population-based, health research specific enterprises have been spearheading health data connectivity and best practices in data governance for the past 30 years [10, 28]. To this point their efforts have been focused largely on enabling secondary uses of individual-level, health-related administrative

data collected by public bodies. The places with longest histories are mainly in Australia (and particularly Western Australia), the UK (in particular Oxford) and Canada (in particular Ontario, Manitoba and British Columbia) [9, 28, 35]. More recently, there have been significant investments by some governments in further developing, linking, and making data available for secondary purposes, including a AUD\$20 million investment by the Australian government in the Population Health Research Network, and £20 million invested in the UK for clinical health data (the Farr Institute) and a further £34 million to expand to social data [13, 46, 60].

Pressures such as the increasing connectivity of data, expanding appetite for access to large data sets, understanding in both the public and private sectors of data as an asset, interest in data well beyond the academic research community, and support for evidence-based policy development, all introduce new challenges for data governance. It is no longer certain that current practices will be fit for purpose for a new era of data. Data governance must be responsive to these new challenges, transparent to those who wish to request access to data, robust in protecting the interests of the individuals and institutions represented in the data, and scalable for expanding sources of data and types of users.

This chapter explores data governance and more specifically governance over access to data, outlining current practice and challenges to that practice, and proposing a framework to support adjudication of secondary uses of data. Given the nature of this volume, we use examples from health care, but the general principles discussed extend easily to other domains as well.

The rest of the chapter proceeds as follows: Sect. 28.2 provides a summary of common data access governance practices for secondary uses of data. Section 28.3 provides an overview of the changes that are occurring in the world of data and data access and how those developments challenge current data access governance. Section 28.4 briefly introduces the concepts of principled and proportionate governance for data access and provides two case studies of those in action. Section 28.5 builds on those case studies and offers an expanded, flexible governance framework, including two worked scenarios. Last, Sect. 28.6 concludes this chapter with some challenges for the future.

28.2 Current Approaches to Data Access Governance

28.2.1 Existing Norms for Data Access Governance

Data governance is about the processes and controls in place to cover the original collection of data, their protection within physical and technical systems, their disclosure and use, and ultimately their archiving or destruction. In other words, there are governance requirements for the full data life cycle [37]. Our focus in this chapter is more specific, on the governance requirements specifically related to organized systems to support secondary uses of data, or the “data access” part of the life cycle.

Where policies and processes for data access exist, most are based on two core elements: (1) privacy legislation, which is built upon the Organization for Economic Cooperation and Development (OECD) privacy principles established in 1980 and reviewed in 2013 [55], and (2) the World Medical Association’s Declaration of Helsinki, a statement of ethical principles for medical research established in 1964 and last revised in 2013 [76], and upon which many more local statements of ethical practices are based (e.g., [7]).

The OECD principles establish fair information principles, essentially outlining what individuals ought to be able to expect when personal data are collected about or from them. In the OECD principles, “personal data” is defined as data about an identified or identifiable individual [55]. The principles include, for example, limiting the collection of information to that which is lawful and where possible with consent of the individual, specifying the purpose of collection, limiting the subsequent disclosure of data, and ensuring that data are well-guarded [55].

Privacy legislation, largely based on principles, similarly pertains to identified or identifiable information; data that are not considered personally identifiable fall outside of the purview of legislation. Legislation also generally includes a clause that is research-enabling, allowing for research access to identifiable information provided (a) the research cannot be done without the data in identifiable form, (b) risks or harms of any proposed data linkage are balanced with the benefits, and (c) the research is in the public interest (for example, FIPPA, 1996).¹

The ethics perspective augments these principles, providing guidance on how human subjects are to be treated in the research process. There is emphasis, for example, on protecting vulnerable populations and balancing any potential direct harm to research subjects with public interest or benefit. In the context of secondary data analyses, ethical guidelines tend to relate to handling of personal data, and limiting researchers’ direct access to personal information. In Canada, for example, the national health research funding agency provided guidance suggesting that data linkage be done not by researchers, but by agencies trusted with that specific part of the research process [4].

The emphasis in current norms is thus on the distinction between what is “personal data”, meaning identified (e.g., with names) or identifiable (e.g., a personal number that could be linked back to an individual), and “non-personal” data, i.e., everything else.

28.2.2 The Preeminence of “Consent or Anonymize” as Approaches to Data Access Governance

The result of this distinction has been a reliance on and preference for individual consent for any current or subsequent use of data [19]. In a consent-based world,

¹http://www.bclaws.ca/Recon/document/ID/freeside/96165_00

uses and authorities and thus data access governance issues, are clearly laid out in advance of data collection. Consent is based on the premise that individuals must fully understand the risks and benefits of engaging in an activity in order to make an informed choice about that engagement [7, 50]. To ensure consent is meaningful, the individual is provided with the necessary information to make an informed decision. What “necessary information” means is open to interpretation, but typically the purpose, collection, use, disclosure, and retention of data must be described in some detail, consistent with the OECD principles [55].

It has been recognized for some time that there are limits to the nature of consent. In the case of secondary use of administrative data, for example, it is accepted that it is not practicable to collect population-based consent on a project-by-project basis, and that the public value of this research may well outweigh any potential harms to individuals stemming from use of their data [4, 50, 56]. This is in part why privacy legislation often carves out research exemptions, albeit in controlled and specific circumstances.

Exemptions in privacy legislation, however, only apply to cases where researchers desire access to personal data. One response to this from a decision-making as well as privacy best practice perspective is to ensure data released are de-identified, therefore technically without legislative or ethical restriction. “De-identification” means personal information is removed, and in some cases data may be aggregated or perhaps even obfuscated in some way [38, 63] but there can remain some risk of re-identification given the right circumstances (and of course intent). The term “anonymous” is generally reserved for data that were either collected without any identifying information in the first place, or data that have been de-identified to the extent that it is impossible to reverse the process [23]. From a regulatory perspective, anonymization and de-identification are considered appropriate because either essentially makes the data a “non-human subject” [1].

Removing information for anonymization generally removes anything that might identify an individual such as cultural, ethnic or disease factors. While protecting privacy, this also eliminates (or at least diminishes) some research possibilities, for example research on often neglected minority groups, individuals with rare diseases, or individuals who live in remote locations [1]. An additional difficulty is that what might be deemed as de-identified or anonymized in the context of a single data set may no longer be the case if those data are linked to other information sources, i.e., if the data are much richer in their content [22]. A reliance on de-identification can lead to a false sense of security, where data custodians feel that they have met their responsibility for data protection, only to find that users develop innovative ways of revealing identities in the data [72].

Underlying all of this is an active debate about at what point data can actually be considered de-identified or anonymous. The identifiability or potential identifiability of data is not truly binary but instead runs along a spectrum such as seen on the horizontal axis of Fig. 28.1, and it is difficult to draw hard lines separating different parts of the spectrum, especially between potentially identifiable and truly anonymous. In practice, nods are made to this challenge by referring to a “reasonableness” or “average person” threshold, or by providing formal assessments

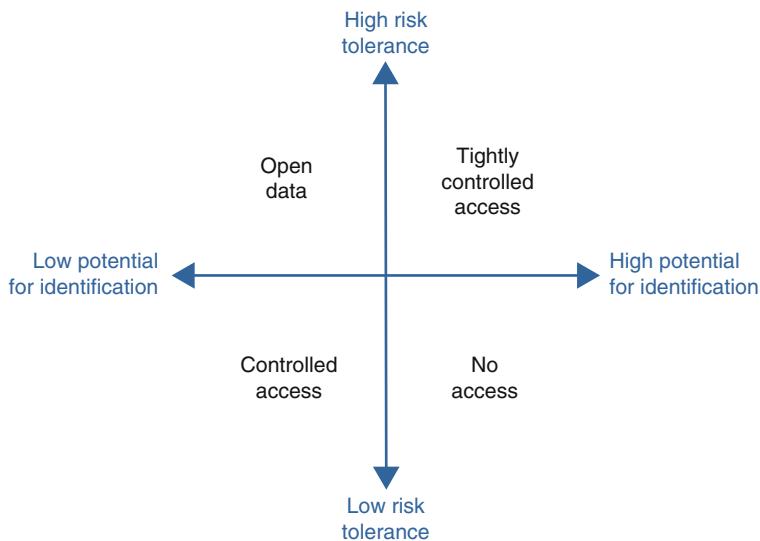


Fig. 28.1 The focus on “identifiability” and “risk” in current data access processes

of the potential for re-identification [15]. Regardless, whether based on human and/or machine input, there must be an assessment that places the data request along a spectrum of potential identifiability, and balances that against the potential good that could come from allowing access.

This is not to dismiss the emerging field of de-identification, perhaps especially in cases where that function is embedded in operational data systems that can control in real time who has access to what level of information under what circumstances [14, 38]. Nor is it to dismiss the value of protecting individuals’ interests and engaging them in the process of understanding how their data are and may be used. The point is simply that there is no silver bullet or one-size-fits-all solution [48]. There are cases where access to more potentially identifiable data—or even fully identified data—is necessary to conduct research that has tremendous potential public value.

Further, it is worth emphasizing, as seen in Fig. 28.1, that the existing focus on either consent or de-identification/anonymization both attempt to control the form of the data to minimize risk. A reliance on de-identification does not represent an approach of balancing needs of different interests, but instead is a risk minimization approach; data are released if the risk is essentially absent. In this scenario, “risk” is not seen as part of an equation, but as something that must be avoided, in which case if risk is present it is the overriding factor of concern.

28.2.3 Existing Data Access Governance in Practice

What emerges without question is that neither a commitment to de-identification, nor a requirement for consent, are substitutes for fully formed data access governance. A governance system for data access provides the balancing required between the public good that can come of use of data with the privacy interests that are embedded in any sensitive information. Legislation and existing norms set up a framework that provides guidance on what to consider for data access governance, but they do not provide a specific blueprint. It is left to individual jurisdictions to set policies and procedures that reflect their local laws, principles and objectives.

There are a number of systems of data access governance that have been established to meet these requirements [20, 36, 43, 57]. These systems are multi-faceted, including multiple inputs and steps. The promotion of the public interest and the protection of individuals' interests are two guiding principles that underpin all of them [26, 41, 54, 70]. Using the example of university-based research, the data access process usually involves: reviews and approvals from granting agencies for approval of the scientific aspects of the research; ethics boards for approval of the treatment of human subjects; and data stewards or guardians for the adjudication of privacy-sensitivity of the data requested, knowledge of the requesters in terms of data use and safeguarding, and the ability of the researcher's environment to provide secure data storage. The specific processes used differ—the forms required, fees charged, physical location of the data and so on—but the basic process has many commonalities. At the same time, these processes are often critiqued because they involve many layers of review, are limited to certain classes of researchers or certain physical locations, take a long time from start to finish, restrict the types of questions that can be addressed and in many cases are not very transparent [39].

One factor related to the critiques of current processes for data access is the default approach of data stewards to their assessment of risk. Referring again to Fig. 28.1, concerns about the potential identifiability of data interact with data stewards' comfort with the implied risk, including both risk to individuals represented in the data and reputational risk to the data provider. Data access governance in practice requires clarity regarding where along the spectrum access will be allowed, and what processes (if any) might help shift either the location of the request along that spectrum or a data steward's tolerance for risk.

There is increasing emphasis in many jurisdictions on providing technical solutions such as safe havens and other secure research environments, where data are stored and accessed regardless of the precise physical location of the researcher [34, 39, 57, 62]. This is in contrast to providing access to data either by giving researchers copies of data on media such as CDs (a practice that is less and less common), or by providing access to detailed data, but only in specific and restricted physical locations [32, 68]. While the latter improves access, the limit to specific physical locations increases time and monetary costs for researchers and therefore lessens the utility of the data [39].

28.3 The Evolution of Data and Implications for Data Access Governance

Against this backdrop are major changes in the world of data that present new challenges to the existing approaches to data access governance. Three interrelated themes we draw out here are big data, open data, and the ubiquity of collection of personal information.

28.3.1 Big Data

The term “Big data” is used to describe data that have extreme volume (many pieces of information taking up lots of computer storage space), variety (different sources and forms, including structure, unstructured, text, media, etc.) and velocity (the speed of data coming in or being accessed) [40]. The sources of data tend to be autonomous and distributed, and the rate of new data creation is extreme, implying that relationships that can be discovered are both complex and evolving [77]. Some claim that 90 % of data that exist today has been created in the past 20 years; whether this is precisely true is less important than the general claim that data, and our collective relationship to them, are changing.

Two developments in the health sphere are particularly significant in the big data realm: the expansion of electronic medical records and the declining cost of gene sequencing. Electronic medical records are significant because they represent an entirely new variety of data compared to administrative data, in that they include free text fields and other unstructured elements. Genetics data on the other hand are significant in their volume; sequencing of genes from a single individual can result in 500 MB of data or more. The cost of sequencing has now decreased to the point that the larger expense is not collecting the data but instead is storing, processing and analyzing those data [52]. Even more importantly, these data will represent a challenge because the most powerful research resulting from them will come from linking different data sources together—combining administrative (including health and social data), electronic medical records, and genomics/epigenomics data [64].

In part because of the rapidly increasing breadth and depth of data, researchers are interested in pursuing analyses that are hypothesis-generating as well as those that are hypothesis-testing. An example of this is the genome-wide association studies commonly undertaken in genomics research [27]. Given an abundance of data and a limited supply of theory, researchers truly are searching for the needle of meaning in a haystack of information. The implication is requests for access to more detailed data for lesser defined research questions, which runs counter to many existing governance structures as well as currently employed privacy tenets.

28.3.2 *Open Data*

The “Open data” movement seeks to make as much information as possible available to the public. Governments increasingly understand data as an asset that has the potential to provide social and economic benefit if they are made available for people to use [18, 24, 44]. This movement comes with significant hype and some early signs of delivering benefit (economic and otherwise) [30], but so far there has been minimal expansion into the health domain.

The OECD, the Research Councils UK and others differentiate the notion of “public interest” from “public good”, with the latter referring to things that are non-rivalrous and non-excluded and available for everyone to engage in a stable and healthy life [53, 61]. To follow on this, the Research Councils UK undertakes that publicly collected data and funded research are public goods, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not infringe intellectual property [61]. From these perspectives, research data and scientific knowledge should be accessible to “families, communities, commercial entities, institutions and governments [in order to] . . . foster the public good” [31].

Efforts towards data liberation are not without concern, however, as a savvier public with access to analytic tools can render de-identified or anonymized data completely identifiable in unanticipated ways [25, 29, 42, 72]. Further, as technologies develop, something that was not identifiable when released could subsequently become identifiable, pointing to the idea that the “reasonableness” standard changes over time. Open data has a different meaning when we are talking about personal information and inherently privacy-sensitive data. There is a big difference between making available a century of weather pattern data and last year’s detailed records of acute inpatient hospital use.

The motivation to make data more available needs to be balanced with recognition that privacy-sensitive data must be governed properly. Systems of governance of data are, arguably, more important than ever. The question will be whether applying the existing rules for data access to these emerging environments will address the needs of all relevant stakeholders.

28.3.3 *The Ubiquity of Collection of Personal Information*

At the same time, the public’s relationship to privacy is changing, in part because we are constantly asked to reveal things about ourselves in exchange for services provided in the digital world. A 2015 editorial in *Science* stated things starkly saying “Privacy as we know it is ending, and we’re only beginning to fathom the consequences” [17]. This is not to suggest the public no longer cares about privacy, but more that with the advent of the internet and smartphones, more personal data than ever is shared with an expanding set of data collectors.

The public willingly exchanges personal information for desired services, including software applications, email, and other on-line services such as those offered through Google. The companies offering these services can legitimately state that they have explicit consent for the collection and use of that information. Whether people truly understand to what they are consenting in these long texts is questionable. In 2014 there was a short experiment undertaken by the Cyber Security Research Institute in London to test the potential dangers of public Wi-Fi use. The researchers offered free Wi-Fi, with a clause in the terms of service stipulating that in exchange individuals would sign over their first-born child [21]. Several people signed up in a short period, though the clause was clearly unenforceable and perhaps the participants understood that.

In a similar vein, a Brooklyn artist offered points to people for various pieces of personal information such as dates of birth, last four digits of a social security number, etc. [58]. The points could then be exchanged for a cookie. The artist posted her terms of use, and they included doing pretty much anything she wished with the data. She had 380 people sign up in the course of an afternoon. This experiment and the Wi-Fi one in London do not necessarily imply that individuals are careless or lack concern about their personal information. Their actions may instead reflect a belief that the people they were dealing with had no ill intent, and/or that in any case their personal information is already “out there” [12].

28.3.4 The Limits of Existing Approaches to Data Access Governance

The world of data itself is changing, including what data are available and how researchers (and others) might interact with them. Governance itself then is likely to change, and arguably must change to meet these changing expectations.

The issue of consent is particularly pointed in the brave new world of data. As discussed, current conceptions of consent require that uses of data must be pre-defined, explicit and time-limited. This is increasingly inconsistent with the aspirations of many who collect data, where future uses are yet-undetermined, and is clearly inconsistent with what the public is willing to accept, at least under some circumstances [45]. As we look to the present and future in areas such as genomics, big data and the rapidly expanding establishment of biobanks for medical research, requiring this up-front and explicit enumeration of all possible uses of data can hobble the value we could gain from them. The real societal and scientific benefits derived from use of data will be maximized if they can be retained for a period of time, and if they are potentially available for new and unanticipated future uses [45].

One proposed solution to this conundrum is to place less emphasis on up-front specific consent and more on controlling users of data. This would include shifting significant accountability to those users, including putting time limits on the storage of personal information [45]. Consistent with this, others have called for modifying

the notion of consent to one of consent to data access governance, which is the processes by which adjudication about access to data would take place [51].

This discussion is particularly important because the debate about potential identifiability of data, and the identification of where data lie on the spectrum depicted in Fig. 28.1, becomes more complex with newer types of data, particularly genomics information. It is understood now that genomics information is inherently identifiable, both for individuals and even family members [42]. Furthermore, there is a growing recognition that data once classified as non-identifiable may become identifiable when linked with other data sets or manipulated by advanced analytics—further blurring the ability to guarantee confidentiality and protection of privacy to individuals. The implication is difficulty in locating a data set on the spectrum of identifiability. The more uncertainty there is about position on the spectrum of identifiability, the less we can rely on de-identification to protect privacy.

Given all of these changes, there is increasing recognition that current approaches to data access need to be modified to reflect a new reality [8, 45]. Current governance arrangements for access to data were largely established prior to the era of the Internet. One response is new or modified legislation, and many jurisdictions are pursuing this option, for example introducing new sector-specific legislation which has the potential to challenge existing norms and practices [59, 69]. More than this, however, is a need to consider more flexible options, approaches that can accommodate uncertainty about identifiability and future uses, expanding potential data users, and more complex research questions.

28.4 A Comprehensive Model for Governance: Proportionate and Principled

A governance model for research data access that is based on defined principles, and where review is proportionate to the level of risk can provide the needed flexible, transparent and solid foundation to address current and future challenges. At its core, such a model involves an approach to governance that is based on both defining how established privacy principles can be incorporated into a research data access framework and how proportionality can be used to create a fair, trustworthy and efficient data access review process. Each of these will be reviewed in turn, starting with proportionality.

28.4.1 *Proportionality*

Proportionality is a general principle of law embedded within the regulatory landscape, in particular European and human rights law, which considers the optimal balance between the level of constraint imposed by a disciplinary measure and the

severity of the proscribed act; it is an attempt to make the punishment fit the crime. Proportionality acts as a criterion for fairness and reasonableness when applied to complex decision-making contexts where interpretative discretion must be used [16] and serves to “... regulate the spaces between hard laws” [41].

In the realm of research data access, proportionality is the practice of ensuring that the review of research data access requests reflects the perceived risk, and the disciplinary mechanisms for non-compliance reflect the damages that may arise. A proportionate approach would adjust the extent and stringency of review according to the potential risk posed by the data request, so that higher risk requests receive more scrutiny. Adopting this type of approach for data access was a key recommendation of the Rawlins Report, a review of the governance and regulation of human health research conducted by the Academy of Medical Sciences at request of the UK government. That report called for regulation that is symmetrical and proportionate, saying “... approving an inappropriate study is clearly unacceptable, but delaying or prohibiting an appropriate study harms future patients as well as society as a whole” [73]. The idea of proportionality is also already embedded in the ethics review process that is often (if not always) a tandem requirement for research data access. As an example, the Canadian Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans [7] states:

Given that research involving humans spans the full spectrum of risk, from minimal to significant, a crucial element of [research ethics board] review is to ensure that the level of scrutiny of a research project is determined by the level of risk it poses to participants ... A reduced level of scrutiny applied to a research project assessed as minimal risk does not imply a lower level of adherence to the core principles. Rather, the intention is to ensure adequate protection of participants is maintained while reducing unnecessary impediments to, and facilitating the progress of, ethical research [7].

Consequently, adopting an approach to requests for access to research data that is symmetric with or proportional to the potential risk of the proposed research depends on understanding the categories or domains that contribute to risk and how the risk assessment maps to the resulting process for adjudication.

28.4.2 Principle-Based Regulation

Principle-based regulation has a long history of application, in both the regulation of financial markets, and the development of human rights and privacy legislation [41, 55, 74]. Principle-based regulation places the emphasis on specifying desired outcomes. One notable example of a principle-based approach in action is the OECD’s Privacy Guidelines, a set of framework principles that address the collection, use, access, accuracy and destruction of personal information [55]. The principles outlined in the Privacy Guidelines have served as a foundation for the development of country-specific privacy legislation, including Canada’s Personal Information Protection and Electronic Documents Act and the European Union Data Protection Directive (95/46/EC). These principles have a long history in data

protection in the medical field, as discussed above, but have traditionally been thought to apply only in cases where researchers are requesting access to personal data.

28.4.3 Case Studies Using Proportionate and Principled Access

There are two known examples of data access processes that include both principles and proportionality. The oldest was the British Columbia Linked Health Dataset (BCLHD) (now part of Population Data BC)² (see Case Study A). A current, more explicit and comprehensive approach was developed by the ScottisH Informatics Programme (SHIP) (see Case Study B) [41, 66]. In the SHIP example, the principles included reflect the values and attitudes expressed by stakeholders through consultations [66]. SHIP proposes a principle-based approach that “... avoids excessive and overly cumbersome procedures [for data access] whilst paying due regard to real risks and seeking appropriate measures where fundamental obligations must be met” [66]. It provides a venue for contemplation of the value and perceived benefit of data use even when proposed activities may conflict with established norms.

Case Study A: The British Columbia Linked Health Database (BCLHD) (1996–2002)³

The British Columbia Linked Health Database, a precursor to Population Data BC, initially brought together core data on health care services utilization and vital events for the province’s residents from 1985 onwards. Adjudication of research data access requests was guided by review of five key issues (domains). The first four had to be met. The fifth determined the potential risk.

- 1. Scientific merit of the research question and method:** Met either by success in a peer review process for grant funded research, or for a student by peer review and approval by his/her supervisory committee.
- 2. Ethical acceptability of the research question and method:** Met with approval from an institutional Research Ethics Board.

(continued)

²We refer to the BC Linked Health Dataset, the precursor to Population Data BC, because it was in this original form that there was an operating proportionate governance model. Population Data BC does not currently operate with such a model.

³The BCLHD transitioned to Population Data BC in 2009. Population Data BC aims to build on BCLHD’s past success and is engaged in efforts to greatly expand its existing data holdings to include educational, occupational, environmental and socioeconomic information. Over time, Population Data BC aims to become the world’s most comprehensive data resource on factors that influence human health, well being and development (<https://www.popdata.bc.ca/>).

3. **Public interest value of the research:** Met with success in peer review from a public funding agency (taken as implying public interest), but also justified by the researcher.
4. **Security of data:** Met by showing care of data would be in accordance with government standards and agreements outlining conditions for data use, return or destruction.
5. **Potential identifiability of data:**
 - (a) No person-specific information required (aggregate data release).
 - (b) Person-specific information but without personal identifiers, detailed birth date or detailed postal code.
 - (c) Person-specific information and potential identifiers (e.g. detailed birth date and/or postal code) but with subsequent contact with individuals and individuals not directly affected by the research.
 - (d) Person-specific information and potential identifiers included in the data, no subsequent contact with individuals but potential for research findings to indirectly affect individuals in the future.
 - (e) Person-specific information and identifiers included in the data with the intent to use information for subsequent contact.

Applications were triaged into two categories for review and approval: fast-track review and full review. Fast track applications were those falling under “potential identifiability of data” categories one and two; all others were subject to full review. In practice, the data access coordinator worked with researchers to move their requests down into category two as much as possible; the vast majority of applications ultimately fell into this category.

Limitations of this model: This model functioned well for many years, but was essentially a risk-minimization approach. It did not consider strategies other than data de-identification for mitigating risk.

Case Study B: The ScottisH Informatics Programme (SHIP) (2012–present)⁴

The ScottisH Informatics Programme is a Scotland-wide research platform that provides research access to linked Electronic Patient Records held by NHS Scotland. Adjudication of research data access requests uses SHIP’s guiding principle and best practices, which are operationalized through four

(continued)

⁴SHIP recently joined the Farr Institute @ Scotland with the intent to expand the expertise, infrastructure and cross-sectoral collaboration for data linkage developed by SHIP (http://www.farrinstitute.org/centre/Scotland/3_About.html).

key benchmarks (domains) each of which includes a spectrum of risk. The domains and optimal (i.e., lowest risk) criteria are:

1. Safe People:

- (a) Everyone in contact with the data is trained in information governance (training materials are vetted).
- (b) Applicant can demonstrate formal track record with administrative data (publications, affiliation with an academic institution, previous application).

2. Safe Environment:

- (a) Level of information security assessed with intended access through a SHIP Safe Haven considered optimal.

3. Safe Data:

- (a) Privacy risk assessment has been undertaken by all projects.
- (b) Data are reviewed in accordance with sensitivity and relation to vulnerable populations.
- (c) Risk of disclosure from linked output file from analysis assessed low (e.g., no full dates, no rare conditions)—high risk (e.g., rare conditions, detailed geographic area).
- (d) General addressing of the privacy and ethical concerns associated with the application.
- (e) Data not used to contact individuals.
- (f) Linkage previously approved for similar purposes.
- (g) Subjects are aware their data may be used for these purposes/permissions are in place for use.

4. Public Interest:

- (a) Applicant demonstrates understanding of context and contribution of the project to the state of public health knowledge.
- (b) Study design and methods noted in application meet objective of the study.
- (c) Public interest not outweighed by commercial interest.
- (d) Formal external peer review undertaken or internal review by an academic institution.

In the SHIP model, a research coordinator reviews the project and assigns a risk score (low, medium, high) for each benchmark. Applications are triaged into four categories based on the total risk score. The four categories are:

Category 0: Public domain data is only requested and no risks are present.

(continued)

Category 1: Risks are minimal and outputs are non-sensitive (e.g., the linkage was previously approved and a safe haven system used for access and storage).

Category 2: Issues are flagged during the review of the application including privacy and security risks and requests for multiple linkages.

Category 3: The application fails considerably to meet an acceptable risk score for at least one of the benchmark.

Applications in Categories 0 and 1 are reviewed only by a research coordinator, Category 2 applications receive fast-tracked advisory board review, and Category 3 applications require full advisory board review.

Limitations of this model: This model is currently limited to researchers affiliated with an academic institution, does not allow for exploratory analyses or hypothesis-generating methods and limits projects to those where public interest outweighs commercial interests.

28.5 Building on the Present: A Flexible, Governance Framework

The contents of this chapter show clearly that data access governance is not a new endeavor, and there are strong foundations on which to build. Nevertheless, changes in data, technology and the climate of data access suggest there is a need to expand our thinking. The aim is for a flexible and adaptable governance framework that incorporates and responds to legislative requirements, long-standing and emerging principles, and builds on input from stakeholders. The framework must be flexible both in being able to respond to further evolution of data and technology, and in being adaptable to local contexts. It is for this reason that we outline a framework, a broad structure that can be modified and applied in different ways by different users.

The proposed framework incorporates six evaluative domains, each of which is associated with four levels of risk ranging from “low” to “very high”. These six domains are outlined here and described in more detail below as well as on Table 28.1:

- **Science:** Scientific merit and potential impact.
- **Approach:** Questions and analyses.
- **Data:** Granularity, sensitivity and justification for the data requested.
- **People:** Experience, affiliation.
- **Environment:** Technical and network infrastructure.
- **Interest:** Public and societal interest.

Table 28.1 A comprehensive framework for proportionate governance: domains for adjunction and associated determination of risk

		Spectrum of risk		
Domain		Low	Medium	High
Science	• Peer reviewed with fundable score		• No peer review in place • Internal review may be in place	Very high
Approach	• Defined questions • Hypothesis testing	• Defined questions with additional exploratory analysis proposed	• Exploratory analysis • Hypothesis generating	• Data mining
Data	• Aggregated data OR • Record-level information but derived variables and/or anonymized information • Anonymization software may be used	• Record-level information • Requested variables tailored to/justified by defined research need	• Record-level, sensitive information and reasonable risk of identifiability • General justification for data use(s) provided; not variable specific	• Record-level sensitive information with the intent to contact individuals
People	• Experienced academic-affiliated user	• New academic users OR • Experienced non-academic users	• Non-academic and non-experienced user	
Environment	• Secure Research Environment (e.g. safe haven)	• Environment with vetted privacy and security controls	• New non-vetted environment	
Interest	• Public interest • No commercial interest	• Public interest • Some commercial interest	• Commercial interest • No short term public interest but potential for longer-term public value	• Commercial interest • No public interest

28.5.1 Science

Science (or scientific merit) considers importance of the question(s) given existing evidence, appropriateness of the proposed methods to answer the question(s), its potential impact and relevance, and the suitability of the scientists to carry out the research. Peer review is an internationally accepted benchmark for ensuring quality and excellence in research [5] and provides an assessment of scientific merit. While peer review is often embedded in the research funding process, funding and the assessment of scientific merit are not always mutually exclusive, as high quality research may not be funded because of the limited availability of funds or the specific priorities of a competition (success rates in granting competitions in the health sector hover around 20 % [6, 47, 49]). As such, a bar of “fundable score” is used.

In practice, an application would be considered low risk if it was reviewed by an external body and received a score that made it eligible for funding, even if funding was not available, and high risk if no peer review took place. If it did not receive a fundable score, it would not meet any conditions for data access, unless or until changes were made to improve scientific merit. A proxy peer review process could be developed for applications without peer review, in which applications are scrutinized according to similar review requirements as standard funding bodies.

28.5.2 Approach

Approach considers the proposed objectives and techniques used to interrogate the data and meet the anticipated outcomes of the research. To be clear, this is not an assessment of methodology, as that would be covered in the “Science” review, but instead speaks to the spectrum of hypothesis-testing (consistent with the prevailing scientific paradigm), to the emerging field of scientific inquiry and predictive analytics, where datasets are interrogated to discover non-obvious patterns and correlations within the dataset, extracting value and generating new hypotheses during the discovery phase [2, 45, 50]. Under current data access regimes, hypothesis-testing research would be more likely to be assessed as low risk, as it presents explicit questions and hypotheses. Hypothesis-generating research, in contrast, with generalized research objectives and assumptions, is often assumed to be of higher risk given its lack of alignment with existing privacy principles of data minimization.

28.5.3 Data

Data considers the sensitivity of the data requested (e.g. amount of data, potential identifiability, vulnerability of study population), extent of linkage among data sources, and the legislative and ethical requirements governing the use of data (e.g., request to contact participants). Data sensitivity also considers the likelihood of a privacy breach and the impact of said breach given the sensitivity and granularity of the data. Low risk requests would be those for aggregate data, while those that wish to receive personal information and contact patients would be deemed very high risk.

28.5.4 People

People represents the human element to the use of data, and includes those who are authorized to access the data and under what conditions. Whereas previous principles focus on appropriate uses and protections, this one looks at the individual or individuals requesting access and places them on a spectrum of risk. This domain is not limited to bona-fide researchers who hold an appointment at an academic institution; it expands the user base to individuals from various organizations who hold portfolios and research interests that might also benefit from access to data (e.g., Health Authorities, clinicians, Non-Governmental Organizations). This approach offers flexibility in considering who is allowed to access the data, and focuses on ensuring that individuals are appropriately qualified and understand their responsibilities for using the data. Baseline qualifications would focus on whether the individual's understanding of legal and ethical issues involved with the data lifecycle meets the proposed research activities and the sensitivity of the data. In practice, an application may be associated with minimal access and personnel risks if personnel have been trained in an accredited Information Governance course; if the applicant has a track record of working with administrative data (publications, affiliation with an academic institution, previous application); and if terms and conditions for data sharing between the personnel and data steward are clearly defined in research agreements.

28.5.5 Environment

Environment refers to the network and technical infrastructure allocated for data handling and storage. Data handling is particularly important as it pertains to linked data, since typically this involves the highest sensitivity of data, and warrants significant levels of scrutiny. Data storage of research extracts may vary from open

data to storage in a pre-vetted environment or safe haven that meets information security measures to approved situations that may operate more on a trust basis.

28.5.6 Interest

Interest refers to the likelihood that the research will result in a tangible outcome that can be widely available to the public [75]. Interest considers the spectrum of potential stakeholders, opinions and benefits to individual and societal interests progressed through the anticipated research. The notion of “public interest”, the act of ensuring that the research would be deemed acceptable by the community at large rather than simply an identifiable individual or sub-group, is considered a cornerstone principle of data use and protection. The difficulty is the potential interpretation that public interests are served only by public entities, which would rule out potential use of data by public-private partnerships, benefit-sharing models or the private sector. A spectrum is again helpful here because while it is recognized that private interests can make use of data that produces public value, the public does express some unease about access to administrative data and the scope of benefit for access requests that are influenced or driven by the commercial sector [11, 65]. Consequently, commercial interests move applications towards the higher risk end of the spectrum.

28.5.7 Translating Risk Assessment to Review Requirements

While this discussion presents the six domains separately, there are clear interconnections among them. For example, understanding the sensitivity of data requested and linkage to other sources will influence requirements for a suitable environment [41]. A request that proposes to use predictive analytics (“Approach”) may be asked to limit the granularity of their request (“Data”).

This points to the fact that understanding how to conduct a risk assessment of a project is only the first step. A review of these domains and identification of the level of potential risk in each provides an overall view, or a heatmap, indicating where risks reside, and what factors are in place to balance or mitigate those risks. There is still a process that needs to be agreed on to translate the level of total risk or “scoring” based on the table to a review process. This would include, for example, determining whether all rows and columns are considered equal, or if instead being “very high risk” on a particular row trumps everything else. This type of risk assessment invites an iterative approach to categorization such as shown in Fig. 28.2, by working with researchers at the onset or upon preliminary review to help them minimize the privacy risks and maximize the public interest value of their research and their skills and competencies as researchers. A project might come in, be rated for a higher level of scrutiny, but then instead of being sent forward

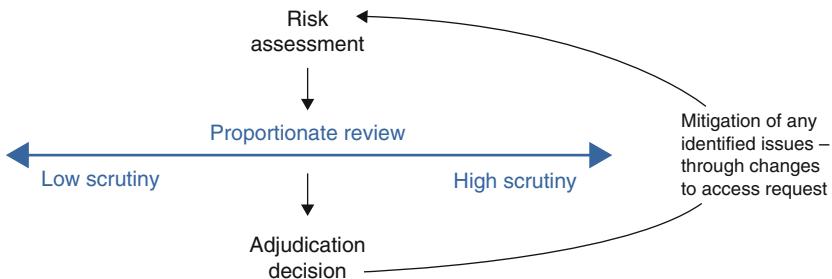


Fig. 28.2 An iterative approach to data access

through the rest of the process, it could be sent back to the researcher to ascertain whether a lower risk rating is possible and desirable. An iterative process mitigates unnecessary burden on researchers and review committees imposed by higher risk applications that could be pre-emptively moved into a lower risk category, and further enables a safe and efficient data access process.

Given the flexibility presented by principle-based frameworks and the jurisdictional variance in legislative and societal requirements, it would be counterproductive for us to present a standalone process for translating the total assessed risk “score” into predetermined review streams. Instead, the next section moves to practice, demonstrating examples of request scenarios and their potential classification along the spectrum of risk for the six domains.

28.5.8 Adjudication Scenarios

The steps shown in Fig. 28.2 are as follows:

1. A request comes in to the coordinating body.
2. The request is assessed using the six domains addressed in the governance framework (Table 28.1) to ascertain the level of requested detail and associated level of risk.
3. Researchers either accept the risk assessment or modify their request to change (presumably lower) it.
4. The risk assessment determines the proportionate review. Those that pose higher risk require a higher level of review, e.g., minimal risk may require no further review whereas high risk may require full review, and may be subject to additional approval conditions, such as access allowed only in a specified data safe haven.
5. Requests may go back to the beginning if the review recommendations prompt researchers to modify their requests to a lower level of risk.

In what follows, we apply the framework to two adjudication scenarios in order to assess the overall risk.

28.5.8.1 Scenario 1

An experienced researcher (“People”) receives grant funding from a public granting agency (“Science”) to conduct a longitudinal cohort study with specific hypotheses about cancer development in women aged 25–50 (“Approach”). The study requests specific de-identified individual-level data fields from the National Cancer Agency, Ministry of Health and Statistics Agency, with annual data refreshes required for 5 years (“Data”). The researcher applies to store and access the research extracts from an approved environment at the National Cancer Research Centre (“Environment”). Both patient groups and the Ministry of Health are part of the research team (“Interest”) (Table 28.2).

Table 28.2 The adjudication scenario 1

	Low	Medium	High	Very High
Science	X			
Approach	X			
Data		X		
People	X			
Environment		X		
Interest	X			

Discussion This scenario is characteristic of a typical application that is routinely reviewed in the data access process in British Columbia through Population Data BC. The study meets external peer review benchmarks, and proposes to test specific questions and defined anticipated outcomes and objectives. Furthermore the researcher is experienced, affiliated with an academic institution with access to a research ethics review board and has published extensively in this area. The risk moves to the right of the spectrum as the data requested is individual-level, will be linked, and is subject to ongoing refreshes without a determined date range. Another medium risk characteristic is the environment where the data will be held; given storage and access will be provided outside of the certified safe haven, the risk moves to the right yet it remains within the lower half of the spectrum as the alternative location has been vetted and is deemed to meet security and data protection requirements.

28.5.8.2 Scenario 2

A new researcher (“People”) proposes an exploratory analysis (“Approach”) of National Pharmaceutical Registry and Ministry of Health data, requesting a large

Table 28.3 The adjudication scenario 2

	Low	Medium	High	Very High
Science			X	
Approach			X	
Data			X	
People		X		
Environment	X			
Interest	X			

set of detailed records (“Data”) to look at individual prescription use and disease development over time. Peer review is not in place (“Science”) however, internal review by the researcher’s affiliated institution has been conducted. The researcher requires significant computing power for the analysis and applies to store and access the research extracts on a vetted safe haven (“Environment”). It is not clear what results might come of this work, though peers express support because of the potential for novel findings (“Interest”) (Table 28.3).

Discussion This scenario is characteristic of a request that reflects the changing data landscape discussed in Sect. 28.3 of this chapter. The request ranks high on the spectrum of risk for half of the domains; it challenges the existing governance and access defaults given it has not received a fundable score through peer review, it sets out to explore the data and generate hypotheses and requests linked, record-level sensitive information that is not tailored to a defined set of questions or time-period. However, the risk is low on other domains given the data will be stored on a vetted safe haven, the research intends to foster additional insights into disease development over time and no commercial or other private interests are at play.

28.6 Conclusion

Changes in technology, the view of data as an asset, commitment to making more data available, the ubiquity of data collection, and growing interest in linking different data sets, all point to a need for transparent governance models for data access. It is no longer feasible to rely on de-identification, anonymization, or consent as primary features of governance. Data are changing and so is the public’s relationship to privacy. It is imperative that we develop systems of governance for data access that reflect public values, have the public’s trust, and encourage accountable uses of data for a spectrum of users.

The proportionate governance framework outlined here offers a comprehensive and flexible framework for data access. It provides a way to meet increasing demand (from both suppliers and users of data), while fully respecting the privacy and other interests of patients and the public. It also enables expanding users beyond the academic community into groups such as non-profits and the private sector and innovators. It would support uses including quality improvement, data mining and

predictive analytics. The framework is flexible enough to meet local circumstance including legal and ethical requirements and rigorous enough to meet the needs of data stewards and the public, who are the individuals represented in the data.

This is not to suggest, however, that implementing a proportionate governance framework will be seamless. Adopting this form of adjudication for data access will be a culture shift. Moving to a proportionate and risk-based approach implies changing from a culture of risk-minimization (or avoidance) to a society that embraces the difficult balancing act between risk and public interest, expanding an ability to make justifiable and appropriate judgments regarding diverse data use and users. Making this shift will require endorsement from key stakeholders, which will only be possible if their needs and concerns are embedded in the definition of principles and the approach to proportionality.

There are many specifics that would need to be determined in order to implement a proportionate governance framework. A critical one is determining the factors that differentiate low, medium, high and very high risk for each review domain. In this chapter we provided suggested starting points, but these need to be adjusted to local circumstance with input from stakeholders, including the public. The framework is an outline that needs filling in. It is a tool for conversations that must lead to an environment in which all stakeholders have their input heard and ultimately give their consent to governance.

There are also decisions that must be made around translating risk assessment to a review process. As with any model of governance, the devil is in the detail. Adopting a proportionate governance framework implies broad commitment to the notion of a more open research culture and creating consensus around important principles. There has been a great deal of talk about the importance of using available data for research and innovation. It is now time to address the relevant principles and how to put them into practice.

Acknowledgements We are grateful to Dawn Mooney for the figures in this chapter, to Megan Engelhardt for help with formatting, and to two anonymous reviewers whose comments greatly improved the content and presentation of the material.

References

1. Anderson, N., Edwards, K.: Building a chain of trust: using policy and practice to enhance trustworthy clinical data discovery and sharing. In: Proceedings of the 2010 Workshop on Governance of Technology, Information and Policies, pp. 15–20. ACM, New York (2010)
2. Biesecker, L.G.: Hypothesis-generating research and predictive medicine. *Genome Res.* **23**(7), 1051–1053 (2013)
3. Brook, E.L., Rosman, D.L., Holman, C.D.: Public good through data linkage: measuring research outputs from the Western Australian data linkage system. *Aust. N. Z. J. Public Health* **32**(1), 19–23 (2008)
4. Canadian Institutes of Health Research.: CIHR best practices for protecting privacy in health research. <http://cihr-irsc.gc.ca/e/46068.html> (2005). Accessed 22 June 2015

5. Canadian Institutes of Health Research.: CIHR peer review manual for grant applications. <http://www.cihr-irsc.gc.ca/e/4656.html> (2009). Accessed 22 June 2015
6. Canadian Institutes of Health Research.: CIHR open operating grant program competitions - frequently asked questions (FAQ). <http://www.cihr-irsc.gc.ca/e/47960.html> (2014). Accessed 22 June 2015
7. Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada.: Tri-council policy statement: ethical conduct for research involving humans. http://www.ncehr-cnerh.org/english/code_2/ (2010). Accessed 22 June 2015
8. Cate, F. H., Cullen, P. & Mayer-Schönberger, V., 2013. Data Protection Principles for the 21st Century, Oxford: Oxford Internet Institute. <http://www.oi.ox.ac.uk/publications/>
9. Chamberlayne, R., Green, B., Barer, M., Hertzman, C., Lawrence, W., Sheps, S.: Creating a population-based linked health database: a new resource for health services research. *Can. J. Public Health (Revue Canadienne De Sant Publique)* **89**(4), 270–273 (1998)
10. Collins, P., Slaughter, P., Roos, N., et al.: Harmonizing research and privacy: Sandars for a collaborative future. Phase II final report: privacy best practices for secondary data use (SDU). http://umanitoba.ca/faculties/health_sciences/Final_Report.pdf (2006). Accessed 22 June 2015
11. Davidson, S., McLean, C., Treanor, S., et al.: Public acceptability of data sharing between the public, private and third sectors for research purposes. <http://www.gov.scot/Publications/2013/10/1304> (2013). Accessed 22 June 2015
12. DenHoed, A.: Give Yourself Away - The New Yorker. <http://www.newyorker.com/culture/culture-desk/giving-away-personal-data-online> (2014). Accessed 22 June 2015
13. Economic and Social Research Council: Big Data Investment: Capital funding - ESRC. http://www.esrc.ac.uk/news-and-events/announcements/25683/Big_Data_Investment_Capital_funding_.aspx (2013). Accessed 22 June 2015
14. Emam, K.E.: Methods for the de-identification of electronic health records for genomic research. *Genome Med.* **3**, 25 (2011)
15. Emam, K.E.: Re-identification risk assessment and anonymization process. <http://www.privacy-analytics.com/de-id-university/white-papers/understanding-risk/> (2013). Accessed 22 June 2015
16. Engle, E.: The history of the general principle of proportionality: an overview. *Dartmouth Law J.* **1**, 11 (2012)
17. Enserink, B., Chin, G.: The end of privacy. *Science* **80**(1), 490–491 (2015)
18. Executive Office of the President.: Memorandum for the Heads of Executive Departments and Agencies 3-9-09. <https://www.whitehouse.gov/the-press-office/memorandum-heads-executive-departments-and-agencies-3-9-09> (2013). Accessed 22 June 2015
19. Faden, R., Beauchamp, T., King, N.: *A History and Theory of Informed Consent*. Oxford University Press, New York (1986)
20. Ford, D., Jones, K., Verplancke, J.P., Lyons, R., John, G., Brown, G., Brooks, C., Thompson, S., Bodger, O., Couch, T., Leake, K.: The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv. Res.* **9** (2009)
21. Fox-Brewster, T.: Londoners give up eldest children in public Wi-Fi security horror show. <http://www.theguardian.com/technology/2014/sep/29/londoners-wi-fi-security-herod-clause> (2014). Accessed 22 June 2015
22. Fullerton, S., Anderson, N., Nicholas, R., Guzauskas, G., Freeman, D., Fryer-Edwards, K.: Meeting the governance challenges of next-generation biorepository research. *Sci. Transl. Med.* **2**(15), 15cm3 (2010)
23. Godard, B., Schmidtke, J., Cassiman, J., Ayme, S.: Data storage and DNA banking for biomedical research: informed consent, confidentiality, quality issues, ownership, return of benefits. A professional perspective. *Eur. J. Hum. Genet.* **11**(2), 88–122 (2003)
24. Government of British Columbia, Ministry of Labour.: Citizens' services and open government: data BC concept of operations. <http://www.gov.bc.ca/citz/> (2012). Accessed 22 June 2015
25. Gymrek, M., McGuire, A., Golan, D., Halperin, E., Erlich, Y.: Identifying personal genomes by surname inference. *Science* **339**(6117), 321–324 (2013)

26. Harmon, S., Graeme, L., Haddow, G.: Identifying personal genomes by surname inference. *Sci. Public Policy* **40**(25), 1–9 (2013)
27. Hirschhorn, J., Daly, M.: Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**(2), 95–108 (2005)
28. Holman, C., Bass, A., Rosman, D., Smith, M., Semmens, J., Glasson, E., Brook, E., Trutwein, B., Rouse, I., Watson, C., de Clerk, N., Stanley, F.: A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. *Aust. Health Rev. (A Publication of the Australian Hospital Association)* **32**(4), 766–777 (2008)
29. Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J., Stephan, D., Nelson, S., Craig, D.: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4**(8), e1000167 (2008)
30. Howard, A.: Open data 500: proof that open data fuels economic activity - techrepublic. <http://www.techrepublic.com/article/open-data-500-proof-that-open-data-fuels-economic-activity/> (2014). Accessed 27 Mar 2015
31. Human Genome Organisation (HUGO), Ethics Committee.: Statement on human genomic databases. *Int. J. Bioeth.* **14**(3–4), 207–210 (2003)
32. Information Governance/Information and Transparency/13630.: Protecting personal health and care information: a consultation on proposals to introduce new regulations - GOV.UK. <https://www.gov.uk/government/consultations/protecting-personal-health-and-care-data> (2014). Accessed 27 Mar 2015
33. Institute on Governance.: Defining governance. <http://iog.ca/defining-governance/> (2015). Accessed 27 Mar 2015
34. Jones, K., Ford, D., Jones, C., Dsilva, R., Thompson, S., Brooks, C., Heaven, M., Thayer, D., Mcnerney, C., Lyons, R.: A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: a privacy-protecting remote access system for health-related research and evaluation. *J. Biomed. Inform.* **50**, 196–204 (2014)
35. Jutte, D., Roos, L., Brownell, M.: Administrative record linkage as a tool for public health research. *Annu. Rev. Public Health* **32**, 91–108 (2011)
36. Kelman, C., Bass, A., Holman, C.: Research use of linked health data – a best practice protocol. *Aust. N. Z. J. Public Health* **26**(3), 251–255 (2002)
37. Khatri, V., Brown, C.: Designing data governance. *Commun. ACM* **53**(1), 148–152 (2010)
38. Kushida, C., Nichols, D., Jadrmicek, R., Miller, R., Walsh, J., Griffin, K.: Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med. Care* **50**, 82–101 (2012)
39. Lane, J., Schur, C.: Balancing access to health data and privacy: a review of the issues and approaches for the future. *Health Serv. Res.* **45**(5), 1456–1467 (2010)
40. Laney, D.: 3D data management: controlling data volume, velocity, and variety. Technical Report, META Group, Stamford. <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf> (2001)
41. Laurie, G., Sethi, N.: Towards principles-based approaches to governance of health-related research using personal data. *Eur. J. Risk Regul.* **4**(1), 43–57 (2013)
42. Lin, Z., Owen, A., Altman, R.: Genomic research and human subject privacy. *Science* **305**(5681), 183–183 (2004)
43. Manitoba Centre for Health Policy – University of Manitoba.: Faculty of medicine – community health sciences – Manitoba centre for health policy. http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departmental_units/mchp/resources/access.html (2015). Accessed 27 Mar 2015
44. Manyika, J., Chui, M., Brown, B.: Big data: the next frontier for innovation, competition, and productivity –McKinsey & Company. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (2011). Accessed 27 Mar 2015
45. Mayer-Schonberger, V., Cukier, K.: Big Data: A Revolution that Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt, Boston (2013)

46. Medical Research Council.: 20 million pounds for new health informatics research institute. <http://www.mrc.ac.uk/news-events/news/20-million-for-new-health-informatics-research-institute/> (2013). Accessed 30 Mar 2015
47. MRC Success Rates.: Medical research council – our research. <http://www.mrc.ac.uk/research/funded-research/success-rates/> (2015). Accessed 27 Mar 2015
48. Narayanan, A.: No silver bullet: de-identification still does not work. <https://freedom-to-tinker.com/blog/randomwalker/no-silver-bullet-de-identification-still-doesnt-work/> (2014). Accessed 27 Mar 2015
49. NIH Success Rates.: NIH research portfolio online reporting tools (RePORT). http://report.nih.gov/success_rates/ (2015). Accessed 27 Mar 2015
50. Nuffield Council on Bioethics.: The collection, linking and use of data in biomedical research and health care: ethical issues. <http://nuffieldbioethics.org/wp-content/> (2015). Accessed 27 Mar 2015
51. O'Doherty, K., Burgess, M., Edwards, K., Gallagher, R., Hawkins, A., Kaye, J., McCaffrey, V., Winickoff, D.: From consent to institutions: designing adaptive governance for genomic biobanks. *Soc. Sci. Med.* **73**(3), 367–374 (2011)
52. O'Driscoll, A., Daugelaite, J., Sleator, R.: "Big data", Hadoop and cloud computing in genomics. *J. Biomed. Inform.* **46**(5), 774–781 (2013)
53. OECD.: OECD principles and guidelines for access to research data from public funding. <http://www.oecd.org/sti/sci-tech/> (2007). Accessed 27 Mar 2015
54. OECD.: OECD guidelines on human biobanks and genetic research databases. <http://www.oecd.org/sti/biotech/> (2009). Accessed 27 Mar 2015
55. OECD.: The 2013 OECD privacy framework. www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf (2013). Accessed 27 Mar 2015
56. OECD.: Strengthening health information infrastructure for health care quality governance: good practices, new opportunities and data privacy protection challenges. <http://www.oecd-ilibrary.org/social-issues-migration-health/> (2013). Accessed 27 Mar 2015
57. Pencarrick-Hertzman, C., Meagher, N., McGrail, K.: Privacy by design at population data BC: a case study describing the technical, administrative, and physical controls for privacy-sensitive secondary use of personal information for research in the public interest. *J. Am. Med. Inform. Assoc.* **20**(1), 25–28 (2012)
58. People Are Willing to give away their personal data for a Cinnamon Cookie. <http://mashable.com/2014/10/01/data-for-cookies/>. Accessed 27 Mar 2015
59. Ploem, M., Essink-Bot, M., Stronks, K.: Proposed EU data protection regulation is a threat to medical research. *Br. Med. J.* **346**, f3534 (2013)
60. Population Health Research Network.: Our funders. <http://www.phrn.org.au/about-us/phrn-funders> (2011). Accessed 30 Mar 2015
61. Research Councils UK.: RCUK common principles on data policy. <http://www.rcuk.ac.uk/research/datapolicy/> (2014). Accessed 6 Mar 2015
62. Roder, D., Fong, K., Brown, M., Zalcberg, J., Wainwright, C.: Realising opportunities for evidence-based cancer service delivery and research: linking cancer registry and administrative data in Australia. *Eur. J. Cancer Care* **23**(6), 721–727 (2014)
63. Rothstein, M.: Is de-identification sufficient to protect health privacy in research? *Am. J. Bioeth.* **10**(9), 3–11 (2010)
64. Schneeweiss, S.: Methods for developing and analyzing clinically rich data for patient-centered outcomes research: an overview. *Pharmacoepidemiol. Drug Saf.* **21**(1), 1–5 (2012)
65. Sethi, N.: Public acceptability of data sharing between the public, private and third sectors for research purposes. <http://www.gov.scot/Publications/2013/10/1304> (2013). Accessed 27 Mar 2015
66. Sethi, N., Laurie, G.: Delivering proportionate governance in the era of eHealth. *Med. Law Int.* **13**(2–3), 168–204 (2013)
67. Stanley, F.: Data for Health. Future Leaders, Sydney (2014)
68. Statistics Canada.: The research data centres program. <http://www.statcan.gc.ca/eng/rdc/index> (2009). Accessed 27 Mar 2015

69. Stenbeck, M., Allebeck, P.: Do the planned changes to European data protection threaten or facilitate important health research? *Eur. J. Public Health* **21**(6), 682–683 (2011)
70. Stevens, L., Laurie, G.: The administrative data research centre Scotland: a scoping report on the legal & ethical issues arising from access & linkage of administrative data. Technical Report ID 2487971, Social Science Research Network (2014)
71. Suissa, S., Henry, D., Caetano, P., Dormuth, C., Ernst, P., Hemmelgarn, B., Lelorier, J., Levy, A., Martens, P., Paterson, M., Platt, R., Sketris, I., Teare, G., Canadian Network for Observational Drug Effect Studies (CNODES): CNODES.: The Canadian network for observational drug effect studies. *Open Med. (A Peer-Reviewed, Independent, Open-Access Journal)* **6**(4), 134–140 (2012)
72. Sweeney, L.: Matching known patients to health records in Washington state data. CoRR [abs/1307.1370](https://arxiv.org/abs/1307.1370) (2013). <http://arxiv.org/abs/1307.1370>
73. The Academy of Medical Sciences.: A new pathway for the regulation and governance of health research. URL <https://www.gov.uk/government/news/a-new-pathway-for-the-regulation-and-governance-of-health-research> (2011). Accessed 27 Mar 2015
74. The Financial Services Authority.: Principles-based regulation: focusing on the outcomes that matter (2007). <https://www.fsa.gov.uk/pubs/other/principles.pdf>. Accessed 27 Mar 2015
75. The Scottish Government.: Joined-up data for better decisions: guiding principles for data linkage. <http://www.gov.scot/Publications/2012/11/9015> (2012). Accessed 27 Mar 2015
76. World Medical Association.: WMA declaration of Helsinki – ethical principles for medical research involving human subjects. <http://www.wma.net/en/30publications/10policies/b3/> (2013). Accessed 27 Mar 2015
77. Wu, X., Zhu, X., Wu, G.D., Ding, W.: Data mining with big data. *IEEE Trans. Knowl. Data Eng.* **26**, 97–107 (2014)

Chapter 29

Epilogue

Aris Gkoulalas-Divanis and Grigoris Loukides

Abstract This chapter provides a summary of the main topics and methods that have been covered in the book, and it draws inferences about various important aspects of medical data privacy. In particular, it discusses issues and techniques related to preserving privacy in: (1) data sharing, (2) distributed and dynamic settings, and (3) emerging applications. Furthermore, it provides an overview of key legal frameworks for the protection of Personal Health Information (PHI) and of techniques required to comply with these frameworks, such as text de-identification and data governance. Moreover, the chapter discusses some promising directions in the field of medical data privacy.

29.1 Introduction

The collected medical data increase in both quantity and complexity, and are becoming an extremely valuable source for analyses that benefit both medical research and practice. However, the privacy of medical data remains an serious concern. On the one hand, privacy breaches continue to occur. For example, more than 600 privacy breaches that are related to healthcare data were reported only in the last 5 years by the U.S. Department of Health and Human Services [11]. Alarmingly, these incidents affect more than 500 and up to 4.9 million individuals each. On the other hand, attacks to individuals' privacy become more sophisticated and new threats are being continuously discovered. To address these issues and preserve data privacy, significant efforts have been put in order to advance technology, law, and policy.

This handbook aimed to provide a comprehensive coverage of the main research areas in medical data privacy. In the remainder of the chapter, we revisit each of

A. Gkoulalas-Divanis (✉)
Smarter Cities Technology Centre, IBM Research, Dublin, Ireland
e-mail: arisdiva@ie.ibm.com

G. Loukides
School of Computer Science & Informatics, Cardiff University, Cardiff, UK
e-mail: g.loukides@cs.cf.ac.uk

these areas and summarize the research challenges and the solutions that have been proposed to offer privacy. In addition, we identify some promising areas for future research in this field.

29.2 Topics and Directions in Privacy Preserving Data Sharing

Protecting the privacy of medical data prior to their sharing is one of the major topics in the field of medical data privacy [4]. One of the pioneering efforts to address privacy in data sharing is the work of Sweeney [9] and Samarati [8]. This work demonstrated that the removal of personal identifiers, such as patient names and phone numbers, is not sufficient to protect the privacy of individuals. This is because other, seemingly innocuous attributes (termed as *quasi-identifiers*) can be used, alone or in combination, to link a patient with their record in the shared dataset. This attack is termed *identity disclosure* and constitutes a serious privacy breach.

Unfortunately, there have been several incidents of medical data publishing, where identity disclosure has transpired. For instance, Sweeney [9] first demonstrated this attack in 2002, by linking a claims database, which contained information of about 135 K patients and was disseminated by the Group Insurance Commission, to the voter list of Cambridge, Massachusetts. The linkage was performed based on patient demographics (date of birth, zip code, and gender) and led to the re-identification of William Weld, then governor of Massachusetts. It was also suggested that more than 87 % of U.S. citizens could be re-identified, based on such (triangulation) attacks. Many other identity disclosure incidents have been reported since then [3]. These include attacks in which (1) students re-identified individuals in the Chicago homicide database by linking the database with the social security death index, (2) an expert witness re-identified most of the individuals represented in a neuroblastoma registry, and (3) a national broadcaster re-identified a patient, who died while taking a drug, by combining the adverse drug event database with public obituaries.

To mitigate the threat of identity disclosure, various data anonymization algorithms have been developed to produce a sanitized counterpart of the original dataset that can be shared with untrusted parties. Chapter 2 surveys the most popular of these algorithms in the context of disseminating patient demographics (relational data) and diagnosis codes (transaction/set-valued data). The chapter explains the objectives of the different methods, as well as the main aspects of their operation. To protect data privacy, anonymization algorithms transform a dataset in order to enforce a formal privacy principle, such as k -anonymity [8, 9], while preserving data utility. There are many algorithms for applying k -anonymity and related privacy principles to various types of data. Some examples of state-of-the-art algorithms, discussed in detail, are those presented in Chaps. 7 and 8.

The effectiveness of anonymization algorithms in terms of offering data utility as well as achieving efficiency, varies significantly. Thus, it is important for a data owner, such as a healthcare institution, to be able to evaluate a range of algorithms and select the best one to apply to a given dataset. Until recently, this was a tedious task, requiring substantial domain knowledge and computer science expertise. However, recent software systems such as those discussed in Chaps. 5 and 6, have simplified the process of producing anonymous datasets by supporting a large number of popular anonymization methods and configurations. In addition, these systems allow, or can be easily extended to allow, the prevention of attacks beyond identity disclosure [5, 6].

Furthermore, there are cases in which protection from a wide class of attacks is necessary. This is possible using the principle of differential privacy [2], which is largely independent of the background knowledge of attackers. Differential privacy ensures that the addition or removal of any record to/from a dataset has little effect on the outcome of a calculation. This is achieved through noise addition (of appropriate magnitude). In several applications that can be supported with the release of aggregate statistics and do not require publishing of (truthful) data at a record level, differential privacy offers a good solution. As discussed in Chap. 3, there are many algorithms for enforcing differential privacy on relational, set-valued, as well as dynamic stream data (e.g., events produced by health monitoring applications). The majority of these algorithms add noise to the data or to the answers of queries that are applied to the data. Minimizing the impact of noise is essential in order to enhance the data utility, but it is not straightforward, particularly for high-dimensional data, as explained in Chap. 4.

While significant progress for enhancing privacy-protection in data sharing has been made, there are still many important research issues that warrant further investigation. First, sophisticated attacks that are possible by combining independently released datasets must be prevented. This should be done under the realistic assumption that coordination between the institutions that have released the datasets is prohibited. An interesting method to prevent a class of such attacks for demographics has been presented in Chap. 8. Further research is needed to prevent these attacks in more complex data, such as data containing both demographics and diagnosis codes [7], or longitudinal data [10]. In addition, it is important to design methods that guard against these attacks, while providing guarantees for the utility of the data in intended query answering and mining tasks. Another class of attacks that need to be thwarted are those performed on aggregated data. These attacks have been the focus of the statistical disclosure control community for years and have led to various methods, which are surveyed in Chap. 9. To increase the adoption of these methods on medical data, it is important to strengthen the protection they offer and to adapt them to operate in web-based data sharing platforms.

Furthermore, to increase the use of privacy-preserving data sharing methods, it is important to address scalability issues. In fact, most of the existing anonymization methods are applicable to datasets that fit into the main memory. Thus, they cannot be used to protect datasets with sizes of several GBs or even TBs. The development of scalable anonymization methods that potentially take advantage of

parallel architectures to solve this problem is therefore worthwhile. Recent steps towards this direction are the works of [12, 13], which are based on the Map/Reduce computing paradigm [1].

29.3 Topics and Directions in Privacy Preservation for Distributed and Dynamic Settings

Medical data are often distributed among multiple parties, such as collaborating healthcare researchers or healthcare organizations that form a consortium. Due to the privacy considerations that these parties typically have, simply sharing raw data among them is not feasible. This has led to the development of privacy-preserving record linkage methods, which allow a set of parties to construct a “global” data view without sharing their raw data. There are two main directions in the development of privacy-preserving record linkage methods, as discussed in Chap. 10. Most of these methods construct the “global” view by employing secure multiparty computation protocols. That is, they perform the operations needed for the record linkage using encrypted data. An alternative way is based on data transformations, and trades-offs privacy for efficiency. That is, each party transforms their data to enforce k -anonymity or differential privacy, and then the record linkage (matching) is performed on the transformed data.

Despite the significant research efforts to develop privacy-preserving record linkage methods, there are some open issues that require further investigation. In terms of privacy, it is important to design flexible methods, which can deal with attributes of different sensitivity as well as with malicious parties. In terms of linked data quality, it is important to design methods that can deal with complex data and offer high accuracy of linkage. An important step towards advancing the research in this field is the establishment of dedicated record linkage centers, as discussed in Chap. 11. The principle behind the operation of these centers is that human intervention can increase the quality and privacy of linked data.

In addition, privacy-preserving methods need to deal with the dynamic aspects of healthcare information management. This is because information is constantly exchanged between different parties, in Health Information Exchange (HIE) systems. These systems are increasingly used for purposes ranging from patient diagnosis and treatment to detection of medical identity theft and financial fraud. However, they pose privacy risks, including potential misuse and unwanted sharing of information, medical identity theft and financial fraud, as explained in Chap. 12. These issues are amplified with the use of large quantities of complex medical data, which are generated by mobile and large-scale networks, and health monitoring medical devices. Tackling these issues while preserving privacy in HIE systems is a promising direction for future research. For example, the security architecture of HIE systems needs to protect data residing in mobile platforms and to account for potentially vulnerable medical devices.

Another important area for preserving the privacy of health information is the design of access control methods. These methods determine what information should be accessed by each party, as well as when, how, and why access to information is performed. To control access to healthcare information in a flexible way, the Role Based Access Control (RBAC) model can be used. This model assigns permissions to users, based on their roles in organizations to denote specific job functions and the associated authorities and responsibilities. RBAC can also serve as basis for solutions that support team collaboration and workflow management, as discussed in Chap. 13. Another way to control access to information is based on patient consent management (i.e., give patients the ability to grant and revoke access to their data). An approach for patient consent management that is designed to deal with the changes to the context of data over time was presented in Chap. 14. Future research directions in the area of consent management include the design of methods that are suitable for mobile devices. In particular, it is interesting to investigate how patients can preserve the control over their data and consent policies, when it is not possible for the system to interact with the device (e.g., after the mobile device has been stolen, lost, or destroyed). Another important direction is to design cross-domain policies for consent management that are transferable between different systems. Such policies can greatly simplify consent management in practice.

Furthermore, there is an increasing interest towards the use of cloud infrastructures for the storage and processing of medical data. This poses certain privacy threats, including malicious data access, intentional data modification, and identity spoofing, which were surveyed in Chap. 15. The majority of existing methods to mitigate these threats are based on cryptographic primitives and are not sufficiently scalable. Promising research directions in this field involve the improvement of the scalability of these methods, as well as the design of techniques to automatically determine whether or not some data must be encrypted in the cloud environment. Another research direction is to improve the auditing and accountability capabilities of methods for managing medical data in the cloud. This is important to warrant the integrity and confidentiality of medical data.

29.4 Topics and Directions in Privacy Preservation for Emerging Applications

The privacy-preserving management, sharing, and analysis of medical data is necessary to support an increasing number of emerging applications. For example, applications featuring genomic, medical image, and Radio Frequency IDentification (RFID) data were discussed in the book, along with applications requiring biomedical signals and health social network data.

Genomic data are essential to realize the vision of personalized medicine (i.e., develop drug and treatments that are tailored for a particular patient). Therefore, they are increasingly shared among healthcare organizations. The sharing, however,

of genomic data poses serious privacy threats, as discussed in Chap. 16. These threats may result in privacy breaches that affect the patient, whose DNA data are shared, as well as the patient's ancestors and descendants. To guard against these threats, it is possible to apply differential privacy [2] to genomic data. An effective algorithm for enforcing differential privacy on genomic data was presented in Chap. 17. The development of differentially private algorithms for genomic data that optimize the utility of the data for a given analysis task, such as performing a statistical association test or building a mining model, is very important. Another approach for preserving the privacy of genomic data is based on cryptography. This approach is applicable when parties are interested in obtaining the result of a particular analysis task, as explained in Chap. 18. Examples of such tasks are genetic association studies and medical tests. To increase the practicality of this approach, it is essential to design methods that are appropriate for the range and complexity of future analysis tasks (e.g., those involving multiple genomes). Furthermore, methods that allow non-expert end-users to run computational genomic tests would be very useful. However, the development of such methods raises questions related to the information that needs to be presented to end-users. These questions include how the presented information can match the expectations of the end-users and the scientific community.

Another type of complex medical data that is becoming increasingly important in applications is images. In fact, medical image data play a central role in healthcare applications related to diagnosis, therapy, follow-up, and clinical research. At the same time, medical image data must be protected from unauthorized access, modification, and copying. This is possible using a combination of encryption and watermarking techniques, as explained in Chap. 19. In addition, it is a legal and ethical requirement to prevent patient re-identification from medical image data and metadata. This requirement is particularly important for neuroimage data and metadata, which can reveal sensitive information about individuals, including serious medical conditions, as it was discussed in Chap. 20. While there are significant efforts for designing frameworks and platforms for the management of neuroimage data, privacy-preserving solutions for such data are lacking. In particular, techniques that are based on objective performance metrics, comply with key regulations, and can be easily integrated in large-scale platforms are needed.

Emerging applications, such as the prevention of infections in hospital environments as well as the monitoring of patients' locations, are based on RFID data. This type of data, however, creates two privacy issues, namely tracking of the location of a patient, and preserving the confidentiality of data that is stored in the RFID tag. Both issues were discussed extensively in Chap. 21. Eliminating tracking is very important, but also challenging. This is because most existing RFID protocols are based on the ability of the reader to retrieve the tag's ID number, which directly allows tracking. On the other hand, preserving the confidentiality of data stored in the RFID tag can be achieved using encryption. The main challenge is to develop scalable methods that are able to deal with the increasing number of RFID tags.

Healthcare applications in hospital environments also feature biomedical signals. A very important type of signal in the context of the diagnosis of heart-related

disorders is the ElectroCardioGram (ECG). More specifically, ECG classification algorithms are central to the diagnosis of ventricular contractions, fibrillation, and tachycardia. However, these algorithms need to be executed on encrypted data, to address privacy considerations, as discussed in Chap. 22. The main challenge here is to preserve both the representation of the ECG data and the intermediate computation results, which are needed by the classification algorithms. Another challenge, which becomes increasingly important, is to be able to execute these algorithms in real-time. Both challenges warrant further research to be fully addressed.

A different type of emerging medical applications are based on health social network data. These applications focus on issues of patient health management, which include providing emotional support, as well as sharing advice and medical information. Health social network data involve personal and sensitive information and need to be protected from attacks, including re-identification, discrimination, profiling and sensitive information disclosure, as explained in Chap. 23. In addition, it is important to protect the communication and interactions between the users of health social networks, which include patients, doctors, and caregivers. Preserving privacy in health social network data is a multifaceted problem, which needs further research. In particular, it is interesting to develop methods that allow the privacy-preserving sharing of health network data, in accordance with user-specified privacy policies and access controls.

29.5 Topics and Directions in Privacy Preservation Through Policy, Data De-identification, and Data Governance

The preservation of medical data is a legal requirement, posed by various legal frameworks. These include laws in the United States, United Kingdom, and Canada, as well as data sharing policies and regulations that have been developed by major research funding organizations in these countries. These frameworks are a first, important step to protecting medical data but may also act as barriers to the global sharing of data, as discussed in Chap. 24. In addition, the breach of privacy laws and data sharing regulations incurs significant financial costs to healthcare organizations. Interestingly, the cause of such privacy breaches is often human error. Therefore, an analysis of the types of human errors is important to improve our understanding of how these errors lead to privacy breaches and how they can be avoided. Such an analysis was performed in Chap. 25, for the case of the HIPAA privacy rule.

To comply with the aforementioned legal frameworks, the shared data must be de-identified (i.e., be devoid from PHI). This is challenging to perform, particularly in the case of unstructured (text) data, such as doctors' clinical notes and prescriptions. Both the detection and the protection of PHI entails significant computational challenges, which are explained in Chap. 26. The detection of PHI can be performed

based on pattern matching (e.g., based on rules provided by domain experts) and machine learning (e.g., classification) algorithms, while privacy protection is typically achieved by the removal or transformation of PHI. An interesting family of transformation-based de-identification methods were reviewed in Chap. 27. These methods replace PHI with surrogates (i.e., synthetic terms that are realistic, in the sense that they can be read naturally) and offer high utility for certain scenarios, such as those involving correspondence between doctors. Despite the progress in de-identification methods, various challenges remain and offer opportunities for future research. These include increasing the generalizability and accuracy of de-identification methods for specific types of PHI, as well as the development of a solid methodology for quantifying the risk of re-identification attacks based on de-identified text data.

Another requirement that is becoming increasingly important to comply with legal privacy frameworks, is medical data governance. This is a procedural requirement, which assists multiple parties (e.g., a consortium of healthcare organizations) to achieve a common goal, such as perform a research study. Chapter 28 provided a comprehensive discussion of medical data governance. To improve the state-of-the-art in this area, more research is needed to deal with the fact that parties often have different privacy and data analysis requirements, as well as with the increased complexity of data. For instance, it is interesting to extend medical data governance solutions to datasets derived from both EHR systems and health social networks.

29.6 Conclusion

Medical data privacy is an area with a broad spectrum of applications, ranging from genomics to patient monitoring and health social networks. As a result, it has attracted significant research interest from the computer science, medical informatics, and statistics communities. This has resulted in various important and practically useful approaches for preserving privacy that were categorized in four areas. The chapter summarized the topics and methods in each of these areas. In addition, it highlighted some possible directions for future work.

References

1. Dean, J., Ghemawat, S.: Mapreduce: a flexible data processing tool. *Commun. ACM* **53**(1), 72–77 (2010)
2. Dwork, C.: Differential privacy. In: ICALP, pp. 1–12 (2006)
3. Emam, K.E., Jonker, E., Arbuckle, L., Malin, B.: A systematic review of re-identification attacks on health data. *PLoS ONE* **6**(12), e28071 (2011)
4. Gkoulalas-Divanis, A., Loukides, G., Sun, J.: Publishing data from electronic health records while preserving privacy: a survey of algorithms. *J. Biomed. Inform.* **50**, 4–19 (2014)

5. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l-diversity. In: ICDE, pp. 106–115 (2007)
6. Nergiz, M.E., Atzori, M., Clifton, C.: Hiding the presence of individuals from shared databases. In: SIGMOD, pp. 665–676 (2007)
7. Poulis, G., Loukides, G., Gkoulalas-Divanis, A., Skiadopoulos, S.: Anonymizing data with relational and transaction attributes. In: ECML/PKDD, pp. 353–369 (2013)
8. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Trans. Knowl. Data Eng. **13**(6), 1010–1027 (2001)
9. Sweeney, L.: K-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl. Based Syst. **10**(5), 557–570 (2002)
10. Tamersoy, A., Loukides, G., Nergiz, M.E., Saygin, Y., Malin, B.: Anonymization of longitudinal electronic medical records. IEEE Trans. Inf. Technol. Biomed. **16**(3), 413–423 (2012)
11. U.S. Department of Health and Human Services.: Breaches affecting 500 or more individuals. <http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrule/index.html> (2015). Accessed 6 Sept 2015
12. Zhang, X., Yang, C., Nepal, S., Chang, L., Dou, W., Chen, J.: A mapreduce based approach of scalable multidimensional anonymization for big data privacy preservation on cloud. In: Cloud and Green Computing, pp. 105–112 (2013)
13. Zhang, X., Liu, C., Nepal, S., Yang, C., Dou, W., Chen, J.: A hybrid approach for scalable sub-tree anonymization over big data using mapreduce on cloud. J. Comput. Syst. Sci. **80**(5), 1008–1020 (2014)

About the Authors

Abstract This section provides contact and biographical information of all 63 authors, who have contributed chapters to this handbook. The authors hold positions in academia or industry, in Europe (France, Germany, Greece, Italy, Luxembourg, Switzerland, and United Kingdom), North America, Asia, Australia and New Zealand. They are presented below in alphabetical order.



Assad Abbas

North Dakota State University

assad.abbas@ndsu.edu

Bio: Assad Abbas completed the Master of Science in Informatics from the University of Skovde, Sweden in 2010. Currently, he is a Ph.D. Candidate in the Department of Electrical and Computer Engineering, North Dakota State University, USA. His research interests are mainly in the areas of cloud computing, e-health, social network analysis, privacy and security of cloud enabled e-health systems, information systems, and ubiquitous computing.



Muhammad Asghar
The University of Auckland
r.asghar@auckland.ac.nz

Bio: Muhammad Rizwan Asghar is a Lecturer at The University of Auckland in New Zealand. Prior to joining this tenure-track faculty position, he was a Post-Doctoral Researcher at international research institutes, including Saarland University in Germany and CREATE-NET in Trento Italy, where he also served as a Researcher. He received his Ph.D. degree from the University of Trento, Italy in 2013. As a part of his Ph.D. programme, he was a Visiting Fellow at the Stanford Research Institute (SRI), California, USA. He obtained his M.Sc. degree in Information Security Technology from the Eindhoven University of Technology (TU/e), The Netherlands in 2009. His research interests include access control, applied cryptography, security, privacy, cloud computing and distributed systems.



Erman Ayday
Bilkent University
erman@cs.bilkent.edu.tr

Bio: Erman Ayday is an Assistant Professor of Computer Science at Bilkent University, Turkey. Before that he was a Post-Doctoral Researcher at EPFL, Switzerland, in the Laboratory for Communications and Applications. He received M.S. and

Ph.D. degrees from the School of Electrical and Computer Engineering, Georgia Institute of Technology, in 2007 and 2011, respectively. He received a B.S. degree in Electrical and Electronics Engineering from the Middle East Technical University, Turkey, in 2005. Erman's research interests include privacy-enhancing technologies (including big data and genomic privacy), wireless network security, game theory for wireless networks, trust and reputation management, and recommender systems. Erman Ayday is the recipient of the 2010 Outstanding Research Award from the Center of Signal and Image Processing at Georgia Tech and of the 2011 ECE Graduate Research Assistant Excellence Award from Georgia Tech.

**Muzammil Baig**

InterSect Alliance International Pty Ltd.

mirza-muzammil.baig@intersectalliance.com

Bio: Mirz Muzammil Baig received the B.Sc. degree in Computer Science (2003) from the University of Central Punjab, Lahore, Pakistan. From 2003 to 2007, Muzammil worked as a research associate at the Al-Khawarzmi Institute of Computer Science in Lahore, Pakistan. He received the M.Sc. degree (2008) from the University of Engineering and Technology (UET) in Lahore, Pakistan and the Ph.D. (2014) from the University of South Australia; both in computer science. In October 2014, he joined InterSect Alliance as a system analyst. His research interests include privacy-preserving data publishing, intrusion detection and system security analysis.

**Mauro Barni**

University of Siena
barni@dii.unisi.it

Bio: Mauro Barni received an M.S. in Electronics Engineering (1991) and a Ph.D. in Information and Communication Engineering (1995), both from the University of Florence, Italy. He is currently with the Department of Information Engineering and Mathematics of the University of Siena, Italy. His research focuses on multimedia and information security, with particular reference to copyright protection, multimedia forensics, and signal processing in the encrypted domain. He has coauthored almost 300 papers and the book “Watermarking Systems Engineering: Enabling Digital Assets Security and other Applications” (CRC Press). From 2010 to 2011, he was the chair of the IEEE Information Forensics and Security Technical Committee of the IEEE Signal Processing Society (SPS). He was appointed a Distinguished Lecturer by the SPS for 2013–2014. He is the editor-in-chief of IEEE Transactions on Information Forensics and Security and is a Fellow of the IEEE and a senior member of EURASIP.

**Maria Bertsimas**

Harokopio University of Athens
mbertsimas@gmail.com

Bio: Maria Bertsimas received her B.Sc. degree in Computer Science from the University of Ioannina, Greece, in 2012. She obtained her M.Sc. in Informatics and Telematics from the School of Digital Technology of the Harokopio University of Athens in 2015. Her expertise is in the field of Management of Information Systems and her current research interests are focused on security and privacy in the context of social networks.

**Luca Bonomi**

Emory University

lbonomi@emory.edu

Bio: Luca Bonomi is a Ph.D. candidate in the Department of Mathematics and Computer Science at Emory University. He received his M.S. and B.S. in Computer Engineering from the University of Padova in Italy. His research interests are in the areas of privacy and security in databases, data mining, and algorithms design.

**Dalel Bouslimi**

Telecom Bretagne

Dalel.Bouslimi@telecom-bretagne.eu

Bio: Dalel Bouslimi received the engineering degree in Computer Science (2009) and the Master degree in Image (2010), both from the Ecole Nationale des Sciences de l'Informatique (ENSI) in Tunisia. She also received the Ph.D. degree in Signal Processing and Telecommunication from the Institute Mines-Telecom-TELECOM Bretagne, France (2013). Since January 2014, she is a post-doctoral researcher in the Information and Image Processing Department of TELECOM Bretagne; LaTIM-Inserm U1101, France. Her research focuses on data protection based on watermarking and cryptographic technologies.

**James H. Boyd**

Curtin University

j.boyd@curtin.edu.au

Bio: James Boyd (B.Sc. (Hons) Mathematics, 1991; Glasgow Caledonian University, Scotland) has been with Curtin University since April 2009. He became Associate Professor and Director of the Centre for Data Linkage (CDL) in May 2010. He was previously employed by the National Health Service in Scotland (NHS) and the Scottish Government. While working with the NHS, he developed and delivered a production record linkage system that still provides monthly updates to NHS Scotland's national linked morbidity and mortality file.

**Rui Chen**

Samsung Research America

rui.chen1@samsung.com

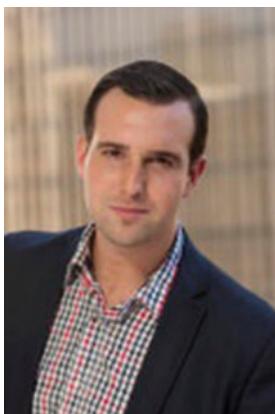
Bio: Rui Chen is a senior research scientist at Samsung Research America. He was a research assistant professor in the Department of Computer Science at Hong Kong Baptist University and a postdoctoral fellow at the University of British Columbia. He received his Ph.D. in Computer Science from Concordia University. His research interests lie in databases and data mining, with a focus on data privacy. He has published over 30 papers in journals and conferences. He has served as program committee co-chair, track chair and member for over ten international conferences and workshops. His research is funded by the Hong Kong General Research Fund.

**Gouenou Coatrieux**

Telecom Bretagne

Gouenou.Coatrieux@telecom-bretagne.eu

Bio: Gouenou Coatrieux is a Professor at the Department of Information and Image Processing, Institute Mines-Telecom, Telecom Bretagne (France) and the head of the Joint laboratory on Security and Processing of Externalized Medical Image Data. He received the Ph.D. degree in Signal Processing and Telecommunication (2002) from the University of Rennes I (France) in collaboration with the Institute Mines-Telecom, Telecom Paris-Tech (France). His primary research interests concern medical information system security and medical data protection by means of watermarking and encryption. He is a member of the International Federation for Medical and Biological Engineering “Global Citizen Safety and Security Working Group” and the European Federation for Medical Informatics “Security, Safety, and Ethics Working Group,” and has contributed to the Technical Committee of “Information Technology for Health” of the IEEE EMBS.

**Edward S. Dove**

University of Edinburgh

edward.dove@ed.ac.uk

Bio: Edward (Ted) Dove is a Ph.D. candidate at the University of Edinburgh School of Law. From 2011 until 2014, he was an Academic Associate at the Centre of

Genomics and Policy at McGill University in Montreal. He holds a B.A. degree in Political Science, and Civil Law and Common Law degrees from McGill, and a Master of Laws degree (LLM) from Columbia Law School in New York City. Ted's primary research interests are in the areas of data privacy regulation, data access and sharing, governance of international research collaboration, genomics public policy, and socio-legal studies of genetic and omics technologies. He is Section Editor at BMC Medical Ethics and an Editorial Board member of OMICS: A Journal of Integrative Biology. He is a member of the International Cancer Genome Consortium's (ICGC) Working Group on Identifiability.

**Liyue Fan**

University of Southern California
liyuefan@usc.edu

Bio: Liyue Fan is a Postdoctoral Research Associate with the Integrated Media Systems Center at the University of Southern California. She received her B.Sc. degree in Mathematics at Zhejiang University in 2008, and Ph.D. in Computer Science and Informatics at Emory University in 2014. Her research work centers around the development of algorithms that facilitate sharing spatiotemporal data without disclosing sensitive information about individual data contributors. She received the Best Student Paper award at DBSec'13 and the departmental Graduate Research Award at Emory in 2014.



Anna M. Ferrante
Curtin University
a.ferrante@curtin.edu.au

Bio: Anna Ferrante has a B.A. (Mathematics) and Ph.D. (Criminology), both from the University of Western Australia (UWA). Between 1996 and 2008, she worked as a Research Fellow and Senior Research Fellow at the Crime Research Centre (CRC) at UWA. During that time she engaged in criminological investigations, including studies of drugs and crime, driving and traffic related crime, and mental illness. While at the CRC, Anna developed the INOIS Record Linkage System. In 2009 she was seconded to Curtin University of Technology to establish the Centre for Data Linkage (CDL). Anna is currently the Deputy Director of the CDL.



Dov Fox
University of San Diego
dovfox@sandiego.edu

Bio: Dov Fox is Associate Professor at the University of San Diego, School of Law, where he teaches and writes in the areas of criminal law and procedure, health law and bioethics, and the regulation of technology. His articles have appeared in leading journals of law and bioethics and he is a regular contributor to the Huffington Post. Dr. Fox received his A.B. from Harvard College, J.D. from Yale Law School, and D.Phil. from Oxford University, where he was a Rhodes Scholar.

**Aris Gkoulalas-Divanis**

IBM Research-Ireland

arisdiva@ie.ibm.com

Bio: Aris Gkoulalas-Divanis received the B.S. from the University of Ioannina (2003), the M.S. from the University of Minnesota (2005) and the Ph.D. from the University of Thessaly (2009), all in Computer Science. His Ph.D. dissertation was awarded the Certificate of Recognition and Honorable Mention in the 2009 ACM SIGKDD Dissertation Award. From 2009 to 2010, he was appointed as a post-doctoral research fellow in the Department of Biomedical Informatics, Vanderbilt University, working on medical data privacy. In 2010, he joined IBM Research in Zurich. Since March 2012, he is working in the Smarter Cities Technology Center of IBM Research in Ireland, leading research in the areas of data privacy and anonymization. Aris has co-authored/co-edited four books in the areas of data anonymization, knowledge hiding, and large-scale data mining.

**Ira Goldstein**

State University of New York at Albany

igoldstein@albany.edu

Bio: Ira Goldstein received his M.B.A. (2004) and his Ph.D. in Information Science (2011) from the University at Albany. He was a Post-Doctoral Research Associate at both the University at Albany and MIT. He led a team that was awarded second place in the 2007 Computational Medicine Center International Challenge:

Classifying Clinical Free Text Using Natural Language Processing. He is currently a Visiting Assistant Professor of Computer Science at Siena College in Loudonville, New York.



Kaitlyn Gutteridge
Population Data BC
kgutteridge@popdata.bc.ca

Bio: Kaitlyn Gutteridge is the Privacy and Governance Lead at Population Data BC. Kaitlyn oversees the organization's multi-faceted data privacy and governance program, and leads negotiation and acquisition efforts for new data holdings. Previously, she trained and worked internationally in built environment and population health research. During her training in epidemiology, she worked with the European Centre on Health of Societies in Transition, a World Health Organization Collaborating Centre, conducting multi-country analyses using administrative and linked data. Kaitlyn holds a B.A. in Health Sciences from Simon Fraser University and a M.Sc. in Public Health from the London School of Hygiene and Tropical Medicine.



John Hale
The University of Tulsa
john-hale@utulsa.edu

Bio: John Hale is a Professor in the Tandy School of Computer Science at The University of Tulsa (TU). He is a founding member of TU's Institute for Information

Security, where he pursues research on attack modeling, cyber trust, and security of cyber physical systems. His research in these areas has been funded by the NSF, DARPA, NSA, and NIJ. He served as Director of TU's Institute for Information Security from 1999 to 2009, overseeing the development of its cybersecurity curricula.

**Peter Hawrylak**

The University of Tulsa

peter-hawrylak@utulsa.edu

Bio: Peter Hawrylak is an Assistant Professor in the Department of Electrical & Computer Engineering and holds a joint-appointment in the Tandy School of Computer Science at the University of Tulsa. He received B.S. Computer Engineering (2002), MSEE (2004), and Ph.D. (2006) degrees from the University of Pittsburgh. His research interests include RFID, security for low-power wireless devices, Internet of Things applications, and digital design. He served as chair of the RFID Experts Group of the Association for Automatic Identification and Mobility in 2012–2013. Peter has over 40 publications and 12 patents. He received the Ted Williams Award from the Association for Automatic Identification & Mobility (AIM) in 2015.



Raymond Heatherly
Vanderbilt University
r.heatherly@vanderbilt.edu

Bio: Raymond Heatherly is a postdoctoral fellow working in the Health Information Privacy Lab at Vanderbilt University, supported by a grant from the National Institutes of Health. He received a B.S. in Computer Science (2004) from Millsaps College in Jackson, Mississippi, and a M.S. (2010) and Ph.D. (2011) in Computer Science from the University of Texas at Dallas.



Dalvin Hill
The University of Tulsa
dalvin-hill@utulsa.edu

Bio: Dalvin Hill received his Ph.D. in Computer Science from The University of Tulsa in 2014. His research areas include medical informatics, information security, and data privacy. His dissertation work led to the development of a novel security framework and blueprint for Health Information Exchanges. Dr. Hill serves on the faculty in the Department of IT and Computer Science at the American University in Afghanistan. He was previously on the faculty of South University in Austin, Texas, and Northeastern State University in Tahlequah, Oklahoma.

**Raquel Hill**

Indiana University

ralhill@indiana.edu

Bio: Raquel Hill is an Associate Professor of Computer Science in the School of Informatics and Computing. Her primary research interests are in the areas of trust and security of distributed computing environments and data privacy, with a specific interest in privacy protection mechanisms for medical-related datasets. Dr. Hill's research is funded by various sources, including the National Science Foundation. She holds B.S. and M.S. degrees in Computer Science from the Georgia Institute of Technology and a Ph.D. in Computer Science from Harvard University.

**Jean-Pierre Hubaux**

École polytechnique fédérale de Lausanne (EPFL)

jean-pierre.hubaux@epfl.ch

Bio: Jean-Pierre Hubaux is a Professor at EPFL. His research aims at laying the foundations and developing the tools to protect privacy in tomorrow's hyper-connected world. He is focusing notably on network privacy and security, with an emphasis on mobile/wireless networks and on data protection, with an emphasis on health-related data and especially genomic data. He has also studied privacy and security mechanisms (especially for mobile networks) in the presence of selfish

players. He held visiting positions at the IBM T.J. Watson and at UC Berkeley. Since 2007, he has been one of the seven commissioners of the Swiss FCC. He is a Fellow of both IEEE (2008) and ACM (2010). After completing his studies in Electrical Engineering at Politecnico di Milano, he worked for 10 years in France with Alcatel, primarily in the area of switching systems architecture and software.

**Xiaoqian Jiang**

University of California, San Diego

x1jiang@ucsd.edu

Bio: Xiaoqian Jiang is an Assistant Professor in the Department of Biomedical Informatics at the University of California—San Diego. He completed his Ph.D. in Computer Science at Carnegie Mellon University. He is an associate editor of BMC Medical Informatics and Decision Making and serves as an editorial board member of the Journal of American Medical Informatics Association. He now works primarily in health data privacy and predictive models in biomedicine. Dr. Jiang won the distinguished paper award from the American Medical Informatics Association Clinical Research Informatics (CRI) Summit in 2013.

**Samee Khan**

North Dakota State University

samee.khan@ndsu.edu

Bio: Samee U. Khan received a B.S. (1999) from Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan, and a Ph.D. (2007) from

the University of Texas, Arlington. He is Associate Professor of Electrical and Computer Engineering at the North Dakota State University. His research interests include optimization, robustness, and security of: cloud, grid, cluster and big data computing, social networks, wired and wireless networks, power systems, smart grids, and optical networks. His work has appeared in over 300 publications. He is on the editorial boards of journals, such as IEEE Access, IEEE Cloud Computing, IEEE Communications Surveys and Tutorials, and IEEE IT Pro. He is a Fellow of IET and BCS.



Florian Kohlmayer

Technische Universität München
florian.kohlmayer@tum.de

Bio: Florian Kohlmayer is a computer scientist and a research assistant at the Chair for Biomedical Informatics at Technische Universität München (TUM), Germany. He holds a Diploma (M.S. equivalent) from TUM. His doctorate studies are supported by a scholarship from the Graduate School of Information Science in Health (GSISH) at TUM and are carried out in close cooperation between the Chair for Biomedical Informatics and the Chair for IT Security. His research interests include IT security and privacy-preserving data management.

**Christopher Kotfila**

State University of New York at Albany
ckotfila@albany.edu

Bio: Christopher Kotfila is a Ph.D. Student in the Informatics program at the University at Albany. His research interests are in natural language processing, machine learning, knowledge organization, and geographic information science. He is a native of Troy N.Y., an avid open source enthusiast, and a hopeless Emacs user.

**Riccardo Lazzeretti**

University of Siena
lazzeretti@diism.unisi.it

Bio: Riccardo Lazzeretti graduated with a degree in computer science & engineering from the University of Siena, Italy, in 2007, where he continued his studies as a Ph.D. student under the supervision of Prof. Mauro Barni in the Information Engineering Department. From November 2009 to May 2010, he was with Philips Lab in Eindhoven, The Netherlands. In 2012, he received a research grant and continued his research in the Information Engineering and Mathematics Department of the University of Siena. His research activity is mainly focused on privacy-preserving applications based on secure two-party computation tools, mainly applied to biomedical, biometric recognition and border control scenarios.

**Xuan Hung Le**

University of Rochester Medical Center
hungfitvn@gmail.com

Bio: Xuan Hung Le received the B.S. degree in Computer Science from Hanoi University, Vietnam, in 2003, and the M.S. and Ph.D. degrees in Computer Engineering from Kyung Hee University, Korea, in 2005 and 2008, respectively. From 2002 to 2003, he served as a software engineer at FPT Corp., Vietnam. From December 2008 to August 2009, he was a Research Professor at the Department of Computer Engineering, Kyung Hee University, Korea. In 2009–2010, he has been a Postdoctoral Research Associate at the Department of Electrical Engineering, University of South Florida. Since October 2010, he was a Post-Doctoral trainee and then Research Assistant Professor at the Biomedical Informatics Research and Development (BIRD) center, University of Rochester Medical Center. His research interests are in wireless sensor networks, ubiquitous computing, biomedical informatics, security, authentication, access control and routing protocols.

**Haoran Li**

Emory University
hli57@emory.edu

Bio: Haoran Li is a Ph.D. student in the Department of Mathematics & Computer Science at Emory University. She is currently working in the AIMS (Assured Information Management and Sharing) group. Her advisor is Prof. Li Xiong. Haoran is currently collaborating with Prof. Xiaoqian Jiang in the iDash Laboratory,

Division of Medicine at the University of California, San Diego. Her research focuses on privacy-preserving machine learning techniques, multi-dimensional data release, and genome data publication under differential privacy.

**Jiuyong Li**

University of South Australia
jiuyong.li@unisa.edu.au

Bio: Jiuyong Li is currently a Professor at the University of South Australia. He received his Ph.D. degree in Computer Science from Griffith University in Australia. His research interests are in data mining, privacy preservation and biomedical informatics. He has published more than 100 papers and his research has been supported by multiple Discovery Grants of the Australian Research Council.

**John Liagouris**

ETH Zurich
liagos@inf.ethz.ch

Bio: John Liagouris is a Post-Doctoral researcher at ETH Zürich, and a member of the Systems Group. Before joining ETHZ, he was a visiting research fellow at the Department of Computer Science, University of Hong Kong (February 2013–July

2014), and a research assistant at the Institute for the Management of Information Systems (IMIS) of the Research and Innovation Center “Athena”, Greece (2009–2015). He obtained a 5-year diploma in Electrical and Computer Engineering in 2008, and a Ph.D. in 2015, both from the National Technical University of Athens, Greece. His main research interests lie in the areas of privacy, graph databases, and data management on heterogeneous hardware.



Divakaran Liginlal
Carnegie Mellon University
liginlal@cmu.edu

Bio: Divakaran Liginlal (Lal) is a Professor of Information Systems in the teaching track at Carnegie Mellon University. Lal holds a B.S. in Telecommunication Engineering from CET, a M.S. in Computer Science from the Indian Institute of Science, and a Ph.D. in MIS from the University of Arizona. Before joining Carnegie Mellon, Lal taught at the University of Wisconsin–Madison’s School of Business, where he was a recipient of the Mabel Chipman Award for Faculty Excellence in Teaching. Lal has also won best teacher awards at CMU’s Qatar campus and the University of Arizona. His research has been published in JMIS, CACM, IEEE-TKDE, IEEE-SMC, EJOR, Computers & Security, and Decision Support Systems. Lal’s research has been supported by Microsoft Corporation, Hewlett Packard, CISCO, Cargill, Qatar Foundation, the Qatar National Research Fund, and the ICAIR at the University of Florida.

**Jixue Liu**

University of South Australia
Jixue.liu@unisa.edu.au

Bio: Jixue Liu got his Ph.D. in Computer Science from the University of South Australia in 2001. His research interests include data transformation and integration, dependency theory in XML and relational data, privacy preservation, and integrity constraint discovery, patterns from sequential data, and text data analysis. Jixue Liu has published in world's top journals in Databases (TODS, JCSS, TKDE, Acta Informatica, etc.)

**Grigoris Loukides**

Cardiff University
g.loukides@cs.cf.ac.uk

Bio: Grigoris Loukides is an Assistant Professor in the School of Computer Science & Informatics at Cardiff University and a Royal Academy of Engineering Research Fellow. His research interests lie broadly in the field of data management with a focus on privacy. His recent research investigates theoretical and practical aspects of data privacy, including algorithmic design, optimization, and formal modeling, and explores applications in healthcare and business. He has received four best paper awards, including an award from the American Medical Informatics Association (AMIA) Annual Symposium, 2009. He obtained a Diploma in Computer Science (2005) from University of Crete, Greece, and a Ph.D. in Computer Science (2009) from Cardiff University, UK.

**Bradley Malin**

Vanderbilt University
b.malin@vanderbilt.edu

Bio: Bradley Malin is an Associate Professor of Biomedical Informatics (in the School of Medicine) and Computer Science (in the School of Engineering) at Vanderbilt University. Since 2008, he has directed the Health Information Privacy Laboratory, which has been sponsored by numerous grants from the National Institutes of Health and National Science Foundation. He is an elected fellow of the American College of Medical Informatics and received the Presidential Early Career Award for Scientists and Engineers in 2009. He received a B.S. in Biological Sciences (2000), M.S. in public policy (2003), and Ph.D. in Computer Science (2006), all from Carnegie Mellon University (USA).

**Kimberlyn McGrail**

University of British Columbia
kim.mcgrail@ubc.ca

Bio: Kimberlyn McGrail, is an Associate Professor at the University of British Columbia, Associate Director of the UBC Centre for Health Services and Policy Research, an Associate with the Centre for Clinical Epidemiology and Evaluation, and a board member and scientific advisor for Population Data BC. Kim's current research interests are in variations in health care services use and outcomes, aging and the use of health care services and understanding health care as a determinant of

health. Kim was the 2009–2010 Commonwealth Fund Harkness Associate in Health Care Policy and Practice. She holds a Ph.D. in Health Care and Epidemiology from the University of British Columbia, and a Master's in Public Health from the University of Michigan.

**Nancy Meagher**

Population Data BC

nancy.meagher@popdata.bc.ca

Bio: Nancy Meagher has been the Executive Director of Population Data BC for the past 8 years. With a Master's in Economics, Nancy has worked internationally in health policy research, hospital management, and pharmaceutical market research, bridging research and its application in policy, business or industry. Her technical experience comes from work as a management consultant, where she led complex, privacy-sensitive development projects for companies including Goldman Sachs. Nancy has grown with Population Data BC, working with staff and key stakeholders to bring it from its inception as the Population Health and Learning Observatory, through integration with the BC Linked Health Database, expansion to include the Population Data Access and Analysis Platform, to its current structure as a single front door for research access: Population Data BC.

**Stephane Meystre**

University of Utah

stephane.meystre@hsc.utah.edu

Bio: Dr. Stephane M. Meystre earned his Ph.D. in Medical Informatics from the University of Utah, his M.D. from the University of Lausanne, Switzerland, and his M.S. in Medical Informatics from the University of California, Davis. He is an Assistant Professor in the University of Utah's Department of Biomedical Informatics. His expertise in clinical informatics research involves the following areas: easing access to clinical data for clinical care and research purposes using advanced techniques such as natural language processing for information extraction and automated de-identification; providing research support by integrating clinical with research data; and integrating research with clinical systems. He also specializes in ontologies development automation, knowledge representation, and clinical text disambiguation. Other areas of interest include: biomedical information and knowledge modeling and representation; telemedicine, teleconsultation, and remote monitoring.

**Noman Mohammed**

University of Manitoba

noman@cs.umanitoba.ca

Bio: Noman Mohammed received a Ph.D. degree in Computer Science from Concordia University in 2012. He is currently an Assistant Professor in the Department of Computer Science at University of Manitoba, Manitoba, Canada. During

2013–2014, he was an NSERC postdoctoral fellow in the School of Computer Science at McGill University, Montreal, Canada. His research interests include private data sharing, privacy-preserving data mining, secure distributed systems, trustworthy cloud computing, and applied cryptography.

**Lucila Ohno-Machado**

University of California San Diego
machado@ucsd.edu

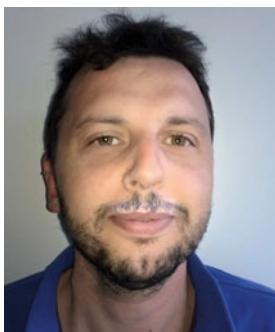
Bio: Lucila Ohno-Machado received her medical degree from the University of São Paulo and her Ph.D. in Medical Information Sciences from Stanford University. She serves as Associate Dean for Informatics and Technology at the UCSD School of Medicine, and chairs the Health System Department of Biomedical Informatics, an informatics research, teaching, and service unit at UCSD. She directs iDASH, an NIH-funded national center for biomedical computing and bioCADDIE, a consortium to develop a biomedical and healthcare data discovery index ecosystem. Her research has been focused on privacy-protecting data sharing, and construction and evaluation of data mining and decision support tools for biomedical research and clinical care. She is PI for the patient-centered Scalable National Network for Effectiveness Research (pSCANNER) project, a clinical data research network funded by PCORI, and the Informatics Core of the Clinical and Translational Research Institute at UCSD.

**Mark Phillips**

McGill University

mark.phillips2@mail.mcgill.ca

Bio: Mark Phillips holds a B.C.L. & LL.B (McGill University) and a Bachelor of Computer Science (University of Manitoba). He has published writing and presented papers on law and technology and on comparative genomic data-privacy law in peer-reviewed law journals, as book chapters, and at conferences. He joined the Centre of Genomics and Policy at McGill University in 2014.

**Giorgos Poulis**

NOKIA Research & Development

poulis2@uop.gr

Bio: Giorgos Poulis received an M.Sc. degree in Techno-economical systems (2004) from the National Technical University of Athens (Greece). He also received a Ph.D. degree in Computer Science (2014) from the University of Peloponnese (Greece), where his research was funded by the Research Funding Program: Heraclitus II. His current research interests include data anonymity for various types of data (relational, transaction, and spatio-temporal), as well as applications in medicine.

**Fabian Prasser**

Technische Universität München
prasser@in.tum.de

Bio: Fabian Prasser is a computer scientist and a research assistant at the Chair for Biomedical Informatics at Technische Universität München (TUM), Germany. He holds a Diploma (M.S. equivalent) and a Ph.D. from TUM. During his doctorate he received a scholarship from the Graduate School of Information Science in Health (GSISH) at TUM and worked as a research assistant at the Chair for Biomedical Informatics and the Chair for Database Systems at TUM. His research interests include graph-structured data, distributed database systems and privacy-preserving data management.

**Sean M. Randall**

Curtin University
sean.randall@curtin.edu.au

Bio: Sean Randall completed his B.Sc. (Hons) in Mathematics and Psychology in 2008 at the University of Sydney. Before joining the Centre for Data Linkage (CDL) in June 2010, Sean worked as a researcher at Macquarie University and then at the Australian Bureau of Statistics. Sean has undertaken record linkage at the CDL, including the linkages for the first PHRN Proof of Concept Project. With colleagues at the CDL, he has conducted research into record linkage methods and published

on various aspects of record linkage, including the effect of data quality on linkage accuracy. More recently, Sean has branched into population-level epidemiological research utilizing linked data.



Panagiotis Rizomiliotis

University of the Aegean
prizomil@aegean.gr

Bio: Panagiotis Rizomiliotis is Assistant Professor at the Department of Information and Communication Systems Engineering of the University of the Aegean, member of the Info-Sec-Lab and external associate at the department of Informatics and Telematics of the Harokopion University. He holds a B.Sc. in Informatics and Telecommunications, an M.Sc. in radioelectrical engineering, and a Ph.D. in Cryptography, all from the Department of Informatics & Telecommunications of the University of Athens. He has received a Marie Curie Fellowship for his postdoctoral research. In 2005 he joined the COSIC research group at Katholieke Universiteit Leuven and worked as a researcher for 2 years. He has published more than 45 international journal and conference papers and book chapters. His research interests include cryptography, coding theory, systems security and information theory.



Giovanni Russello
The University of Auckland
g.russello@auckland.ac.nz

Bio: Giovanni Russello is a Senior Lecturer and Leader of the Digital Security Programme at the University of Auckland, New Zealand. He received his Ph.D. from the Eindhoven University of Technology, The Netherlands. After obtaining his Ph.D., he was a Research Associate in the Department of Computing at Imperial College London, UK. His research interests include policy-based security systems, privacy and confidentiality in cloud computing, smartphone security and applied cryptography.



Sarowar Sattar
University of South Australia
sarowar@gmail.com

Bio: Sarowar Sattar received the B.Sc. degree in Computer Science and Engineering (2004) from the Rajshahi University of Engineering & Technology (Bangladesh), the M.Sc. degree in Information Technology (2010) from the Royal Institute of Technology (Sweden) and the Ph.D. degree in Computer Science (2015) from the

University of South Australia (Australia). From 2004 to 2008, Sarowar worked as a Lecturer in the Computer Science and Engineering department of the Rajshahi University of Engineering and Technology. In August 2008, Sarowar was promoted to an Assistant Professor at the same university.



Nakeisha Schimke

The University of Tulsa

nakeisha.schimke@utulsa.edu

Bio: Nakeisha Schimke received her B.S. in Computer Science (2003), M.S. in Computer Science (2005), and Ph.D. in Computer Science all from The University of Tulsa (2011). Her research areas include information assurance, operating system security, medical informatics, neuroinformatics and data privacy. Dr. Schimke developed a security and privacy framework integrated into a neuroimage archival platform, and pioneered practical techniques for privacy preservation in neuroimage data sets.



Natalie Shlomo

University of Manchester

natalie.shlomo@manchester.ac.uk

Bio: Natalie Shlomo received a Ph.D. in Statistics (2007) from the Hebrew University (Jerusalem). In 2007 she joined the University of Southampton in the UK and is now a Professor in Social Statistics, School of Social Sciences at the University

of Manchester, UK. Before, she worked at the Israel Central Bureau of Statistics from 1981 to 2007. Her areas of interest are survey methodology and official statistics, with emphasis on survey design and estimation, non-response analysis and adjustments, record linkage, statistical disclosure control, statistical data editing and imputation, and small area estimation. She is an elected member of the International Statistical Institute, an elected Council member for the International Association of Survey Statisticians and a fellow of the Royal Statistical Society.



Spiros Skiadopoulos
University of Peloponnese
spiros@uop.gr

Bio: Spiros Skiadopoulos is currently an Associate Professor at the Department of Informatics and Telecommunications at University of Peloponnese. He received a diploma and a Ph.D. degree from the National Technical University of Athens and a M.Phil. degree from UMIST, UK. He has worked in a variety of areas including data management, knowledge representation and reasoning. His current research interests include anonymity, and big data management and processing (e.g., very large graphs, string and time series data). He has published more than 40 papers in leading journals and conferences. Moreover, he have served in the program committee of several venues and participated in various research and development projects.

**Amber Stubbs**

Simmons College

amber.stubbs@simmons.edu

Bio: Amber Stubbs received her Ph.D. in Computer Science from Brandeis University. Her dissertation research involved creating an annotation methodology to extract high-level information—such as medical diagnoses—from narrative texts. With James Pustejovsky, she co-authored the book “Natural Language Annotation for Machine Learning” (O’Reilly, 2012). Her interests include natural language processing, corpus annotation, and interpreting clinical narratives. She is currently an Assistant Professor at Simmons College in Boston, MA.

**Peter Szolovits**

Massachusetts Institute of Technology

psz@mit.edu

Bio: Peter Szolovits is a Professor of Computer Science and Engineering in the MIT Department of Electrical Engineering and Computer Science, where he is the head of the Clinical Decision-Making Group within the MIT Computer Science & Artificial Intelligence Laboratory. His research centers on the application of AI

methods to problems of medical decision making, natural language processing to extract meaningful data from clinical narratives to support translational medicine, and the design of information systems for health care institutions and patients. He is the 2013 recipient of the Morris F. Collen Award of Excellence from the American College of Medical Informatics.



Qiang Tang

University of Luxembourg

qiang.tang@uni.lu

Bio: Qiang Tang is a postdoctoral researcher in the APSIA group, SnT, University of Luxembourg. His research interests include applied cryptography and privacy-preserving data engineering. Currently, he is a principal investigator in the Junior CORE project, BRAIDS—Boosting Security and Efficiency in Recommender Systems, funded by the Luxembourgish government. He received his B.S. (1999) from Yantai University and M.S. (2002) from Peking University in China. He received his Ph.D. degree in Information Security and Cryptography from Royal Holloway, University of London in 2007. Before moving to Luxembourg, he was a postdoctoral researcher at University of Twente, the Netherlands.

**Manolis Terrovitis**

Institute for the Management of Information Systems (IMIS)

mter@imis.athena-innovation.gr

Bio: Manolis Terrovitis is a researcher at the Institute for the Management of Information Systems of Research Center “Athena”. His research interests include privacy preservation, data anonymization and Big data analytics. He received his Ph.D. from the National Technical University of Athens (2007) and has been with the Department of Computer Science of The University of Hong Kong as a post-doctoral researcher (2007–2008). He has published works on data privacy and data analytics in many international conferences and journals in data management (e.g., VLDB, VLDBJ, TKDE). Dr. Terrovitis has been involved in several national and EU projects in the area of data management and analytics.

**Christos Tryfonopoulos**

University of Peloponnese

trifon@uop.gr

Bio: Christos Tryfonopoulos received his B.Sc. degree from the University of Crete (2000), and his M.Sc. (2002) and Ph.D. (2006) from the Technical University of Crete. In 2006 he joined the Databases and Information Systems Department at Max-Planck Institute for Informatics (MPII) in Germany as a post-doctoral researcher, and from 2007 to 2008 he was leading the P2P and Information Management research group at MPII. In 2009 he joined the Department of Computer Science & Technology, University of Peloponnese as a Lecturer and a

member of the Software & Database Systems Lab. Since 2013 he is an Assistant Professor at the Department of Informatics and Telecommunications, University of Peloponnese. His research interests include data and information management, large-scale distributed systems, and digital libraries.



Özlem Uzuner

State University of New York
ouzuner@albany.edu

Bio: Özlem Uzuner is an Associate Professor at the State University of New York, Albany and a research affiliate at MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). She received her Ph.D. from MIT. Her dissertation was on Identifying Expression Fingerprints Using Linguistic Information for the purposes of copyright infringement detection, even when documents are paraphrased. She studies applications of Natural Language Processing to real-world problems. Her current projects include information retrieval, medical language processing, information extraction, and de-identification.

**Iraklis Varlamis**

Harokopio University of Athens
varlamis@hua.gr

Bio: Iraklis Varlamis is an Assistant Professor at the Department of Informatics and Telematics of Harokopio University of Athens. He obtained a Ph.D. in Computer Science from Athens University of Economics and Business of Greece in 2003. His research interests vary from data-mining and the use of semantics in web mining to graph analysis in social networks. He has published several articles in international journals and conferences in the areas of web document clustering, the use of semantics in web link analysis and web usage mining, word sense disambiguation using thesauruses, and more.

**Joseph Walker**

MyHealth
joe.walker@myhealthaccess.net

Bio: Joseph (Joe) Walker is Privacy Officer for the MyHealth Access Network, a Health Information Exchange in Oklahoma, and directs its development and ongoing operations. He led MyHealth's Privacy and Security committee in its development of a comprehensive privacy and security framework for the organi-

zation and its stakeholders to share and use protected health information legally, securely, and appropriately. Mr. Walker has also developed a comprehensive return-on-investment calculator for HIEs. He has 9 years previous experience with Sprint Nextel's Network Operations division, as well as experience in project engagements with GE Wind, C.R. Bard, Integrated Care Management, Children's Healthcare of Atlanta, and Multicell Packaging Inc.



Dongwen Wang

University of Rochester Medical Center
dongwen_wang@urmc.rochester.edu

Bio: Dongwen Wang is Associate Professor of Biostatistics and Computational Biology, Associate Professor of Medical Informatics, and Director of Biomedical Informatics Research & Development Center at the University of Rochester Medical Center. Dr. Wang's research is on the use of information science and technology to improve biomedical research, clinical education, patient care, and public health. The specific focusing areas of his research include modeling and representation of biomedical knowledge in computer-interpretable format, management of biomedical data in specific context of workflow and team collaboration, development and dissemination of online resources, and delivery of technology-mediated behavioral interventions to improve clinical processes and healthcare outcomes. Dr. Wang serves as domain expert for grant review, paper review, organization of academic conferences, and development of technical standards. His research is funded by the federal government, New York State, and other sponsors.

**Shuang Wang**

University of California San Diego
shw070@ucsd.edu

Bio: Dr. Shuang Wang is a postdoctoral researcher at the Department of Biomedical Information at University of California San Diego. He received his Ph.D. degree in Electrical and Computer Engineering from the University of Oklahoma. His background is in data privacy, data compression and machine learning. His current research interests focus on the medical and genome data privacy, genomic data compression, and GPU-based high performance computing.

**Li Xiong**

Emory University
lxiong@emory.edu

Bio: Li Xiong is an Associate Professor in the Department of Mathematics and Computer Science and the Department of Biomedical Informatics at Emory University where she directs the Assured Information Management and Sharing (AIMS) research group. She holds a Ph.D. from Georgia Institute of Technology, an M.S. from Johns Hopkins University, and a B.S. from University of Science and Technology of China, all in Computer Science. She also worked as a software engineer in IT industry for several years prior to pursuing her doctorate. Her

areas of research are in data privacy and security, distributed and spatio-temporal data management, and health informatics. She is a recent recipient of the Career Enhancement Fellowship by Woodrow Wilson Foundation, a Cisco Research Award, and an IBM Faculty Innovation Award. Her current research is supported by NSF, AFOSR, and PCORI.

Glossary

AAD Average Absolute Distance.

ABAC Attribute Based Access Control.

ABE Attribute Based Encryption.

ACA Patient Protection and Affordable Care Act.

ACP Access Control Policy.

ACTORS automatic Creation and lifecycle managemenT Of authoRization policieS.

ADNI Alzheimer's Disease Neuroimaging Interface.

ADVS Attribute-based Designated Verifier Signature scheme.

AES Advanced Encryption Standard.

AHPA Additive Homomorphic Proxy Aggregation.

AIHW Australian Institute of Health and Welfare.

AMC Albany Medical Center.

ANOVA ANalysis Of VAriance.

APC Artial Premature Contractions.

APEC Asia-Pacific Economic Cooperation (privacy framework).

API Application Programming Interface.

AQIM Angle Quantization Index Modulation.

ARE Average Relative Error (measure).

BA Budget Absorption (approach).

BAM Binary sequence Alignment/Map (file).

BAT Battery Assisted Passive (tag).

BCLHD British Columbia Linked Health Dataset.

BD Budget Distribution (approach).

BMI Body Mass Index.

BoB VHA Best-of-Breed clinical text de-identification system.

BP Branching Program.

BPP Bits (of message) Per Pixel (of image).

BVP Brain Volume Preservation.

CADPP Critical Assessment of Data Privacy and Protection (community).

CAT Cornell Anonymization Toolkit.

CBA Clustering-Based Anonymizer (algorithm).

CBC Cipher Block Chaining.

CCP-CABE Constant-size Ciphertext Policy Comparative Attribute-Based Encryption.

CDL Centre for Data Linkage (Australia).

CDR Clinical Data Repository.

CDSS Clinical Decision Support System.

CECUP Clinical Education Center for Upstate Providers.

CEI (New York State HIV) Clinical Education Initiative.

CEW Commutative Encryption and Watermarking.

CFB Cipher FeedBack.

CHeReL Centre for Health Record Linkage (Australia).

CHIR VHA Consortium for Healthcare Informatics Research.

CI Certified Institution.

CKE Content Key Encryption.

CM Classification Metric.

CNV Copy-Number Variation.

CP Cloud Platform.

CP-ABE Ciphertext Policy Attribute Based Encryption.

CRF Conditional Random Field.

CSCW Computer-Supported Cooperative Work.

CSP Cloud Service Provider.

CSV Comma-Separated Values.

DAC Data Access Committee.

DAC Discretionary Access Control.

DAC Data Access Committee.

dbGaP database of Genotypes and Phenotypes.

DC-QIM Distortion Compensated Quantization Index Modulation.

DCT Discrete Cosine Transform.

DES Data Encryption Standard.

DFT Discrete Fourier Transform.

DICOM Digital Imaging and Communications in Medicine.

DIN Drug Identification Number.

DIS DISassociation (algorithm).

DLB Data Linkage Branch (Western Australia).

DM Discernibility Metric.

DM Dither Modulation.

DMP Dossier Medical Personnel.

DNA Deoxyribonucleic Acid.

DOS Denial of Service (attack).

DP Differential Privacy.

DP Data Provider.

DPA Data Protection Authority.

DPD Data Protection Directive.

DS4P Data Segmentation For Privacy (initiative).

DSS Data Sharing Service.

DTC Direct To Customer (service).

DTE Distribution-Transforming Encoder.

DUA Data Use Agreement.

DUC Data Use Certification.

DVS Designated Verifier Signatures.

EC Evaluation Center.

ECB Electronic CodeBook.

ECG ElectroCardioGram.

ECMC Erie County Medical Center.

EE Expansion Embedding.

EEA European Economic Area.

EEG Electroencephalography.

EFPA Enhanced Discrete Fourier Transform.

EFW Encryption Followed by Watermarking.

EHR Electronic Health Record.

EMR Electronic Medical Record.

ENISA European Network and Information Security Agency.

EPC Electronic Product Code.

ePHI electronic Protected Health Information.

ESPAC Efficient and Secure Patient-centric Access control Scheme.

FFR Facial Feature Removal.

FHE Fully Homomorphic Encryption.

FIDIS Future of Identity in the Information Society.

FIPPA Freedom of Information and Protection of Privacy Act (Canada).

FIPS Fair Information Practice Principles.

FTC Federal Trade Commission.

GC Generalization Cost (measure).

GCP Global Certainty Penalty (method).

GDS Genomic Data Sharing (policy).

GP General Practitioner.

GPPM Genomic Privacy-Preserving Mechanism.

GPRD General Practice Research Datalink (UK).

GUI Graphical User Interface.

GWAS Genome-Wide Association Study.

HA Health Authority.

HBC Honest-But-Curious (adversary).

HD Hellinger Distance (measure).

HE Honey Encryption (framework).

HF High Frequency.

HGP Human Genome Project.

HGS Human Genomic Study.

HHS (Department of) Health and Human Services.

HIDE Health Information DE-identification.

HIE Health Information Exchange.

HIMSS Healthcare Information and Management Systems Society.

HIPAA U.S. Health Insurance Portability and Accountability Act.

HIS Health Information System.

HITECH U.S. Health Information Technology for Economic and Clinical Health Act.

HL7 Health Level 7 (standard).

HME Homomorphic Encryption.

HPC Health Personal Card.

HPE Hierarchical Predicate Encryption.

HS Histogram Shifting.

HSN Healthcare Social Network.

IBE Identity-Based Encryption.

IC Institutes and Centers.

ICT International Classification of Diseases.

iDASH integrating Data for Analysis, anonymization and SHaring.

ILM Information Loss Metric.

IOM Institute of Medicine.

IoT Internet of Things.

IRB Institutional Review Board.

ISM Industrial, Scientific and Medical (frequency band).

IT Information Technology.

JPEG Joint Photographic Experts Group (standard).

JWD Joint Watermarking-Decryption.

KHS Key Holder Site.

KP-ABE Key Policy-Attribute Based Encryption.

K-S Kolmogorov-Smirnov (test).

LBP Linear Branching Program.

LCE Lossless Compression Embedding.

LDF Linear Discriminant Function.

LIMS Laboratory Information Management System.

LONI Laboratory for NeuroImaging.

LRU Least Recently Used (eviction policy).

LSB Least Significant Bit.

LSH Locality-Sensitive Hashing (function).

LUT Look-Up Table.

MAC Mandatory Access Control.

MCHP Manitoba Centre for Health Policy (Canada).

MCP Multi-Cloud Policy.

MD Multinomial-Dirichlet synthesizer.

MD5 Message Digest 5 (algorithm).

MeDS Medical De-identification System.

MHC Mental Health Center.

MIST MITRE corporation Identification Scrubber Toolkit.

MK Masking and Key manager.

MLE Maximum Likelihood Estimation.

MLI Minimum Linkage Information.

MPI Master Patient Index.

MRE Matching Relative Error.

MRI Magnetic Resonance Imaging.

MRSE Multi-keyword Ranked Search over Encrypted Data.

MSB Most Significant Bit.

MU Medical Unit.

NCBI National Center for Biotechnology Information.

NCP Normalized Certainty Penalty (measure).

NHS National Health Service.

NIH National Institutes of Health.

NIST National Institute of Standards and Technology.

NLP Natural Language Processing.

NN Nearest Neighbour.

NN Neural Network.

NPO (Swedish) National Patient Summary (initiative).

NRMSE Normalized Root-Mean-Square Error.

NSERC Natural Sciences and Engineering Research Council of Canada.

NSR Normal Sinus Rhythm.

OASIS Organization for the Advancement of Structured Information Standards.

OCBE Oblivious Commitment Based Envelop (protocol).

OCR Optical Character Recognition.

OECD Organization for Economic Cooperation and Development.

OFB Output FeedBack.

ONC Office of the National Coordinator (in HHS).

ONS Office for National Statistics (UK).

ORLS Oxford Record Linkage Study.

OSN Open Social Network.

OT Oblivious Transfer.

PACS Picture Archiving and Communications Systems.

PAP Policy Administration Point.

PBE Password Based Encryption.

PBR Principles-Based Regulation.

PCE Patient Controlled Encryption.

- PCEHR** (Australian) Personally Controlled Electronic Health Record (initiative).
- PDP** Policy Decision Point.
- PDS** Privacy-preserving Disease Susceptibility test.
- PEP** Policy Enforcement Point.
- PET** Privacy-Enhancing Technology.
- PGP** Personal Genome Project.
- PHDA** Priority-based Health Data Aggregation.
- PheWAS** Phenome-Wide Association Study.
- PHI** Protected Health Information.
- PHR** Personal Health Record.
- PII** Personally Identifying Information.
- PIM** Private Information Matching (model).
- PINQ** Privacy INtegrated Queries.
- PIP** Policy Information Point.
- PIPEDA** Personal Information Protection and Electronic Documents Act (Canada).
- PKE** Public Key Encryption.
- PPGL** Privacy Preserving Group Linkage.
- PPRL** Privacy Preserving Record Linkage.
- P-QIM** Perpetual Quantization Index Modulation.
- PRAM** Post-RANDomization (method).
- PRE** Proxy Re-Encryption.
- PSD** Private Spatial Decomposition.
- PSNR** Peak Signal-to-Noise Ratio.
- PSUC** Prevention and Substance Use Center.
- PU** Processing Unit.
- PVC** Premature Ventricular Contractions.
- QA** Quality Assessment.
- QDF** Quadratic Discriminant Function.
- QID** Quasi-IDentifier/Quasi-IDentifying attribute.

QIM Quantization Index Modulation.

RBAC Role Based Access Control.

RDBMS Relational DataBase Management Systems.

RDM Rational Dither Modulation.

RE Relative Error.

RF Radio Frequency.

RFID Radio Frequency IDentification.

RLS Record Locator Service.

ROI Region Of Interest.

RONI Region Of Non-Interest.

RRC Resource and Referral Center.

RT Relational-Transaction.

RTLS Real-Time Location System.

SA Situation Awareness.

SAIL Secure Anonymized Information Linkage.

SAM Sequence Alignment/Map (file).

SAMHSA U.S. Department of Substance Abuse and Mental Health Services Administration.

SARS Severe Acute Respiratory Syndrome.

SDL Statistical Disclosure Limitation.

SHA-1 Secure Hashing Algorithm 1.

SHE Somewhat Homomorphic Encryption.

SHIP ScottisH Informatics Programme.

SKE Symmetric Key Encryption.

SLK Statistical Linkage Key.

SMC/SMPC Secure Multi-Party Computation (protocol).

SNB Slide Negative Binomial (distribution).

SNOMED-CT Systematized Nomenclature of Medicine-Clinical Terms.

SNP Single Nucleotide Polymorphism.

SNR Signal-to-Noise Ratio.

SNV Single Nucleotide Variation.

SPED Signal Processing in the Encrypted Domain.

SPIN Shared Pathology Informatics Network.

SPU Storage and Processing Unit.

SSE Searchable Symmetric Encryption.

SSH Secure SHell (protocol).

SSHRC Social Sciences and Humanities Research Council of Canada.

SSL Secure Sockets Layer (protocol).

SSN Social Security Number.

SVM Support Vector Machine.

SVT SupraVentricular Tachycardia.

SWRL Semantic Web Rule Language.

SWT Standard Widget Toolkit.

TA Trusted Authority.

TC Technology Center.

TF-IDF Term Frequency-Inverse Document Frequency.

TLE Two Layer Encryption.

TLS Transport Layer Security.

TPDC Testing, Post-Exposure Prophylaxis, and Diagnosis Center.

TR Teleo-Reactive (programming).

UCSP Uniform Ciphertext by Swapping Plaintext (approach).

UCUP Uniform Ciphertext/Uniform Plaintext (approach).

UHF Ultra-High Frequency (band).

UID Unique IDentifier.

UL Utility Loss (function).

UML Unified Modeling Language.

UMLS Unified Medical Language System (US National Library of Medicine).

UNICON UNIversal Constraint ONtology.

URMC University of Rochester Medical Center.

VF Ventricular Fibrillation.

VHA U.S. Veterans Healthcare Administration.

VSA Valuematic Scaling Attack.

VT Ventricular Tachycardia.

WBAN Wireless Body Area Network.

WFE Watermarking Followed by Encryption.

WGS Whole Genome Sequencing.

WHO World Health Organization.

WorldII World Legal Information Institute.

XACML eXtensible Access Control Markup Language.

XML Extensible Markup Language.

XNAT eXtensible Neuroimaging Archive Toolkit.

ZIP U.S. Zone Improvement Plan (postal codes).

Index

A

ABAC, 315
About the Authors, 775
access control, 302, 399
 patient-controlled encryption, 399
access policy, 330
 encoding, 330
 interpretation, 333
ACTORS approach, 371
AES, 499, 500
anonymization algorithm, 21, 25, 27, 95
 k -member, 25
 agglomerative, 25
 Anatomize, 28
 Apriori, 27, 100
 CBA, 27, 164
 CBFS, 25
 Cluster, 99
 COAT, 100
 DiffGen, 47
 DiffPart, 49
 disassociation (DIS), 165
 DPCopula, 42
 DPCube, 44
 EFPA, 46
 Filtering/Sampling, 44
 Fourier transform, 43
 full subtree bottom-up, 99
 genetic, 25
 genomic data anonymization, 449
 greedy, 25, 28
 heuristic strategies, 24
 Hilb, 25, 28
 iDist, 25, 28
 Incognito, 25, 28, 99

Infogain Mondrian, 25
KACA, 25
LPA, 43
LRA, 27, 100
LSD Mondrian, 25
MDAV, 25
mHgHs, 27
Mondrian, 25, 99
NNG, 25
NoiseFirst, 44
OLA, 25
P-HP, 47
PCTA, 100
PrivBasis, 50
Privelet+, 43
PSD, 45
RBAT, 28
recursive partition, 27
risk-based anonymization, 118
sample-based, 28
SDL methods, 207
StructureFirst, 44
SuppressControl, 28
TDCControl, 28
TDS, 25
top-down, 99
tree-based, 28
UAR, 27
UGACLIP, 27
VPA, 27, 100
anonymization tools, 86
 μ -Argus, 143
 ARX, 86, 113
 CATtool, 86, 142
 deid software, 703

anonymization tools (*cont.*)

- HIDE, 144, 704
 - MITRE identification scrubber toolkit, 703
 - PINQ, 144
 - sdcMicro, 143
 - SECRETA, 143
 - TIAMAT, 86, 143
 - UTD tool, 142
 - VHA BoB de-id system, 706
- APEC, 363
- API, 123
- ARE measure, 40, 90, 96, 104
- attribute disclosure, 19, 20, 27, 115
- audit, 304, 399
- availability, 496

B

blocking, 257

C

- Canadian Privacy legislation, 658
- CEI collaboration, 327, 328
- CEI responsibilities, 317
- chi-squared test, 69
- clinical education initiative, 314, 316
- cloud computing
 - adversarial models, 399
 - authenticity, 410
 - collusion resistance, 406
 - data integrity, 402
 - deployment models, 393
 - opportunities, 391
 - patient confidentiality, 400
 - privacy requirements, 397
 - privacy threats, 394
 - service provider, 390
 - unlinkability, 412
- composition attack, 181
- composition theorem, 39
- confidentiality, 496
- consent management, 300, 375, 432, 740
 - data access governance, 740
 - emergencies, 368
 - essential elements, 365
 - legal framework, 363
 - limitations, 366
- cryptography, 402, 497
 - asymmetric cryptosystems, 499
 - block cipher algorithms, 500
 - encryption mechanisms, 498
 - symmetric cryptosystems, 499

D

- DAC, 315
- data anonymization, 25, 26, 62
- data curation, 303
- data encryption, 239, 482
 - honey encryption, 482
- data governance, 271, 294, 739
 - approaches, 739
 - case studies, 748
 - challenges, 744
 - current systems, 743
 - framework, 752
 - limits of approaches, 746
- data ownership, 297
- data privacy breaches, 688
- data protection authorities, 654
- data tampering, 396
- data transformation, 18, 22, 152
 - n*-grams, 244, 246
 - constraint-based generalization, 95
 - embedding, 244, 248
 - full-domain generalization, 94
 - generalization, 187, 244, 245
 - hierarchy-based generalization, 95
 - micro-aggregation, 207
 - multidimensional generalization, 94
 - noise addition, 210
 - phonetic encoding, 244, 250
 - PRAM, 207, 209
 - random rounding, 217
 - rank swapping, 207
 - record swapping, 207, 217
 - set-based generalization, 157
 - stochastic perturbation, 218
 - subtree generalization, 94
 - suppression, 244, 245
 - utility-constrained approach, 152
- data utility, 18, 23, 119, 131, 160, 162, 211, 214, 449
 - AAD measure, 211
 - ARE measure, 90
 - average equiv. class size, 119
 - average relative error (ARE), 23
 - classification metric (CM), 23
 - discernability metric (DM), 23, 119
 - frequency-based utility constraints, 163
 - generalization cost(GC), 23
 - hierarchy-based utility constraints, 162
 - information loss metric (ILM), 23
 - LM, 23
 - Loss, 119
 - MRE measure, 160
 - NCP, 96
 - non-uniform entropy, 119

- normalized certainty penalty(NCP), 23
RCV measure, 212
similarity-based utility constraints, 163
utility loss (UL), 96, 97
- dbGaP, 426
 data access, 428
 data submission, 428
- dedication, v
- defacing MRI, 536
- deid software tool, 703
- denial of service, 396
- DES, 499
- DICOM, 499, 529
- differential privacy, 35, 64, 197, 219, 243, 448
 ϵ -differential privacy, 220
 applications, 66
 composability theorem, 65
 evaluation, 67
 histogram, 35
 randomization, 185
 sensitivity, 65, 448
 survey sampling, 219
 w-event privacy, 53
- DiffPart, 49
- disclosure risk, 213, 215
 differencing, 213
 group attribute, 213
 Hellinger's distance, 215
 individual attribute, 213
 table linking, 213
- DNA, 430, 431, 463, 474
- domain ontology, 319
- E**
- ECG classification
 autoregressive modeling, 572
 complexity analysis, 587
 ECG segment classification, 574
 feature extraction, 573
 garbled circuits, 578
 hybrid protocols, 579
 neural networks, 590
 plain protocol, 572
 privacy-preserving diagnosis, 584
 privacy-preserving hybrid classifier, 597
 privacy-preserving quality evaluation, 599
 quadratic discriminant function, 573
 quantization error analysis, 585
- EnCoRe, 368
- ENDORSE, 368
- enhanced permission set, 319
- equivalence class, 188
- EU Data Protection Directive, 234, 364, 652, 699
- exponential mechanism, 38
- F**
- facial reconstruction, 540
- FAST framework, 52
- FTC, 363
- G**
- generalization, 18, 187, 244, 245
- GenoGuard protocol, 483
- genome privacy, 444, 463, 465
 ethical principles, 432
 GenoGuard protocol, 483
 genomic data anonymization, 449
 GPPM framework, 485
 Homer's attack, 435
 Human Genome Project, 444
 kin genomic privacy, 465
 personalized medicine, 472
 privacy analysis, 453
 privacy attacks, 445, 463, 467
 privacy risks, 434, 463
 privacy-protection techniques, 446
 private use in research, 477
 protection technologies, 434, 470
 SAM file, 470
 utility criteria, 449
 weak passwords, 481
 Whole Genome Sequencing, 444
- genome-wide association studies (GWAS), 426, 427, 478
- GWAS data sharing policy, 427, 430
- geometric mechanism, 37
- GINA, 431
- glossary, 815
- H**
- health information exchange
 architecture, 293
 privacy issues, 295
 record locator service, 293
- health social networks, 618
 account deletion, 624
 communication tracking, 626
 current HSNs, 631
 de-anonymization attacks, 625
 digital dossier aggregation, 623
 identity theft, 626
 inference attacks, 625

health social networks (*cont.*)

- information leakage, 627
- phishing, 626
- privacy issues, 618
- privacy principles, 621
- privacy requirements, 627
- privacy threats, 622, 623
- secondary data collection, 624

HIDE, 704

HIPAA Limited Dataset, 663

HIPAA Privacy Act

- PHI categories, 700

HIPAA Privacy Rule, 234, 294, 390, 431, 528, 659, 680, 699

IBM Ponemon study cost of privacy breach, 681

PHI categories, 726

privacy breaches, 688

HIPAA Safe Harbor, 662

HIPAA Security Rule, 680

HITECH Act, 294, 390, 660, 680, 698

Homer's attack, 435

homomorphic encryption, 236, 252, 407, 570, 576

I

IBM Ponemon study, 681

ICT, 570

iDASH, 437

identifiability, 742

identity disclosure, 19, 20, 115, 150

information loss measures, 211

integrity, 496

J

journalist scenario, 64

JWE approach, 516

K

KL-divergence, 196

Kolmogorov–Smirnov test, 69

L

Laplace mechanism, 37

linear branching programs, 571, 580

linear discriminant function, 571

logistic regression, 69, 74

M

m-Health, 306

MAC, 315

membership disclosure, 115

microdata, 203

minimum linkage information, 279

MITRE identification scrubber toolkit, 703

MRE measure, 160

MRI defacer, 536

MRI images, 529

N

neuroimages, 529

ADNI, 543

BVP, 541

de-identification, 535

facial recognition, 532

facial reconstruction, 540

FFR, 541

LONI, 543

MRI defacer, 536

privacy in archives, 543

privacy risks, 530

privacy threat scenarios, 530

skull stripping, 536

volume rendering, 532

NIH, 426

NLP, 698

noise addition, 210

non-parametric methods, 41, 43

DiffGen, 47

DPCube, 44

EFPA, 46

Filtering/Sampling, 44

Fourier transform, 43

LPA, 43

NoiseFirst, 44

P-HP, 47

Privelet+, 43

PSD, 45

StructureFirst, 44

non-repudiation, 497

NPO initiative, 698

O

OASIS standard, 362

oblivious transfer, 576, 577

OECD, 363, 648, 650

open data, 744

P

PACS, 494

Paillier, 499, 576

parametric methods, 41

- PCEHR, 698
PHI, 294, 389, 528, 680, 703, 722, 723
 PHI categories, 722, 723
PII, 367
policy, 372, 373
 authorization, 372
 decision point, 372
 enforcement, 372
 template, 374
policy-based authorization, 362
priority-based data aggregation, 406
privacy legislation, 645
 accountability approach, 646
 adequacy approach, 646
 Canadian data sharing policies, 666
 Canadian Privacy legislation, 658
 cost of privacy breach, 682
 data privacy legal frameworks, 642
 data sharing policies, 664
 de-identification, 646
DPA, 654
EU Data Protection Directive, 652, 699
GWAS Policy, 665
HIPAA de-identification fields, 662
HIPAA Limited Dataset, 663
HIPAA Privacy Rule, 659, 699
HIPAA Safe Harbor, 662
HITECH Act, 660
pseudonymisation, 646
UK Data Protection Act, 656
use of PHI for research, 660
Wellcome Trust, 669
privacy model, 18, 19, 36, 117, 155
 (d, α) -linkable model, 189
 (h, k, p) -coherence, 20
 δ -presence, 118
 ℓ -diversity, 20, 118
 ρ -uncertainty, 21
 k -anonymity, 19, 62, 117, 239
 k -map anonymity, 20
 k^m -anonymity, 155
 p -sensitive k -anonymity, 20
 t -closeness, 20, 118
 (a, k) -anonymity, 20
 differential privacy, 36, 37
 encryption/hashing, 239
 PS-rule anonymity, 20
privacy risk, 426
privacy threat, 19
privacy-preserving record linkage, 235, 279, 304
 n -grams, 246
 blocking, 236, 257
 clustering, 259
embedding, 248
homomorphic encryption, 252
hybrid approaches, 256
mapping, 259
matching, 236, 238
minimum linkage information, 279
phonetic encoding, 250
PPRL model, 236
secure multi-party computation, 251
sorted neighborhood, 258
private-FIM algorithms, 50
 PrivBasis, 50
privilege escalation, 396
proportionality, 747
prosecutor scenario, 64
- Q**
quasi-identifier, 19, 84
- R**
randomization, 185
RBAC, 315, 319
 access delegation, 319
 collaboration constraint, 319
 implementation, 329
 separation of duty, 319
 temporal constraint, 319
 universal constraints, 319
RC4, 501
record linkage, 233, 268
 centralized model, 275
 data governance, 271
 duplicate detection, 233
 entity resolution, 233
 information security, 274
 international centres, 268
 privacy challenges, 271
 research infrastructure, 268
 separated models, 276
record linkage centres, 268
record linkage quality, 282
regression analysis, 212
repudiation, 396
RFID, 550, 562
 device management, 557
 frequency bands, 551
 passive HR tags, 552
 passive UHF tags, 552
 privacy issues, 555
 tags, 550, 551
 technological benefits, 556
risk analysis, 113, 118, 138, 204

RSA, 499
 RT-dataset, 84, 96

S

s.p.e.d. technology, 570
 SAM file, 470
 secure multi-party computation, 251, 570
 secure multiparty computation, 243
 semi-honest security model, 576
 semi-parametric methods, 41, 42
 DPCopula, 42
 sensitive attribute, 19
 sensitivity, 37, 65, 198, 448
 situation awareness, 685
 privacy breaches, 686
 skull stripping, 536
 SNPs, 463, 474, 485
 SNR, 599
 soundex encoding, 250
 spoofing, 396
 staircase mechanism, 38
 statistical disclosure limitation, 201
 suppression, 18, 244, 245
 synthetic data, 35

T

telemedicine, 306, 497
 applications, 497
 teleo-reactive policies, 369
 evaluation, 370
 representation, 369
 text de-identification, 724, 725
 clinical text anonymization, 708
 definition, 698
 deid tool, 703
 HIDE tool, 704
 measures, 712
 methods, 701
 MIST, 705

MITRE scrubber toolkit, 703
 PHI categories, 700, 722, 723
 PHI detection, 701
 PHI removal, 701
 Physionet deid, 704
 surrogate generation, 724, 725
 VHA BoB de-id system, 706
 traceability, 497

U

UK Data Protection Act, 656
 UMLS metathesaurus, 703
 utility policy, 160

V

VAERS database, 528
 VHA BoB de-id system, 706

W

w-event privacy, 53
 watermarking, 503
 basic properties, 505
 followed by encryption, 513
 general scheme, 503
 index authenticity control, 503
 index integrity control, 503
 lossless watermarking, 509
 lossy modulations, 507
 medical images, 506
 metadata insertion, 503
 non-reversible, 513
 quantization index modulation, 508
 reversible, 513
 Wellcome Trust, 669

X

XACML, 362